GRADO EN ESTADÍSTICA

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

TRABAJO FIN DE GRADO:

EXPLOTACIÓN Y MODELOS PARA PREDICCIÓN DE DATOS DE UN CENTRO EDUCATIVO

Trabajo Fin de Grado realizado por:
Francisco José Barrera Gómez
[francisbarrera@gmail.com]

· *junio de 2016* ·

Profesor Tutor:
Antonio Beato Moreno

GRADO EN ESTADÍSTICA

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

TRABAJO FIN DE GRADO:

EXPLOTACIÓN Y MODELOS PARA PREDICCIÓN DE DATOS DE UN CENTRO EDUCATIVO

Trabajo Fin de Grado realizado por Francisco José Barrera Gómez

· *junio de 2016* ·

Profesor Tutor: Alumno:

Fdo: **Antonio Beato Moreno** Fdo: **Francisco José Barrera Gómez**

Agradecimientos

A mi familia por su paciencia en mis horas de dedicación.

A mi tutor por su interés y transmisión de conocimientos.

junio de 2016

Resumen

El presente documento constituye la memoria del Trabajo Fin de Grado en Estadística por la Facultad de Matemáticas de la Universidad de Sevilla y su título es "Explotación y modelos para predicción de datos de un centro educativo".

El trabajo está organizado en varios capítulos que se detallan a continuación:

- En el capítulo uno a modo de introducción, se contextualiza cuál es la situación y características del centro educativo referenciado mostrando brevemente la plataforma de formación online sobre la que se trabaja, esta es, Moodle en su versión 2.5.9. Se concluye con los objetivos propuestos que se pretenden conseguir.
- El segundo capítulo está dedicado a la explotación descriptiva de datos. Para ello, previamente se detalla la estructura general de los mismos, el proceso de obtención y concretamente cuáles son los datos disponibles con los que vamos a trabajar. A continuación, se realiza un análisis estadístico descriptivo inicial con el fin de conocer y visualizar la distribución de los mismos. La última parte de este capítulo se centra en detallar el diseño y construcción de un informe automático de estadísticas descriptivas para agilizar la obtención de resultados según unos requisitos establecidos.
- En el tercer capítulo abordaremos la parte dedicada a los modelos de predicción para dar respuesta a los dos problemas que en principio se han planteado: predicción de alumnos que pueden causar baja en su matrícula con un mes de antelación y estimación de la calificación en la convocatoria ordinaria de mayo/junio con los datos relativos hasta la segunda evaluación.
- Seguidamente, en el cuarto capítulo se exponen las conclusiones obtenidas con este proyecto, detallando las aportaciones al centro. En el último apartado de este capítulo, se reflejan algunas de las líneas futuras de trabajo.
- En el quinto y último capítulo, se detallan a modo de anexos, los códigos fuente de los scripts en lenguaje R construidos para implementar las técnicas y metodologías propuestas.

Por último, se indican las referencias bibliográficas y documentación online consultados.

Abstract

This document is the end-of-degree project in Statistics at the Faculty of Mathematics of the University of Seville. Its title is "Data Exploitation and Forecasting Models for a school".

This project is organized as follows:

- Chapter 1 is an introduction. It presents the situation and specific characteristics of the mentioned school, showing the online education platform (Moodle) version 2.5.9 under usage. It ends with the objectives planned.
- Chapter 2 is devoted to the descriptive analysis of the data. Firstly, their general structure is described. Secondly, the collection process and specifically the exact data under analysis are presented. Next, an initial descriptive analysis is made in order to know and visualize their distribution. The last section of this chapter is focused on explaining in detail the design and implementation of an automatic report of descriptive statistics which eases the data collection according to some pre-established requirements.
- Chapter 3 is about the proposed forecasting models to solve the two challenges initially posed: i) Forecasting one month in advance the amount of students un-enrolling, ii) Estimating the grade of a student at the end of the third term of the course given the grades of the first two terms.
- Chapter 4 establishes the conclusions of this project, detailing the contributions. Potential future research lines are also identified.
- Chapter 5 – Annexes include the R scripts that implement the techniques and methodologies proposed.

Finally, the bibliography is shown in the References.

Índice general

Lista de tablas

Lista de figuras

Capítulo 1

Introducción

Las características de los sistemas de educación presencial y a distancia presentan ventajas según las circunstancias del alumnado. Los sistemas de educación en modalidad semipresencial se suelen incluir en los modelos de educación a distancia. En la enseñanza presencial el profesor dispone de la interacción (observación) física diaria con el alumno como recurso para tomar decisiones durante el proceso de evaluación.

Aunque desde el punto de vista del alumno aspectos como la comodidad de no tener que desplazarse al centro educativo, organizarse el tiempo de estudio de una forma más flexible o una buena calidad de los materiales, son aspectos valorados positivamente, para el profesor implica una mayor planificación a largo plazo, buenos conocimientos de las Tecnologías de la Información y la Comunicación para la elaboración de materiales en formato digital y la dificultad de tener un seguimiento continuo del alumno.

El estudio e implementación práctica llevadas a cabo en este Trabajo Fin de Grado, en adelante TFG, se han hecho en un centro educativo público andaluz en el que se imparten las enseñanzas de ESA (Enseñanza Secundaria Obligatoria para Adultos) y Bachillerato para adultos en la modalidad no presencial, también llamada enseñanza a distancia (online).

La realidad ha sido que tras los resultados obtenidos fruto de este proyecto, en el centro se ha abierto la puerta a la consideración o inclusión del tratamiento y análisis estadístico de datos como herramienta necesaria e imprescindible para el conocimiento de la actividad educativa diaria(Álvaro Jiménez Galindo 2010).

El hecho de tratarse de un proyecto innovador en el centro ha condicionado en parte la metodología seguida, concretamente el no tener disponibles los datos de las calificaciones de la tercera evaluación hasta junio han retrasado la última parte del trabajo referente a regresión lineal. Es necesario por tanto, para el curso que viene, fijar un calendario de generación de datos brutos fuente para obtener resultados en el tiempo de una forma más eficiente.

Aunque el objetivo último del proyecto ha sido explotar los datos disponibles para obtener resultados a partir del establecimiento a priori de varios requisitos aplicando diversas técnicas estadísticas, gran parte del tiempo y el trabajo empleados se ha centrado en preparar los datos para poder tratarlos en este caso con el lenguaje de programación R, crear informes en varios formatos (HTML, PDF) haciendo uso además de LATEX y RMarkdown. Ello nos avisa

de que en la práctica el análisis estadístico de datos requiere, además de los conocimientos puros de Estadística, de herramientas y técnicas informáticas de tratamiento de datos.

1.1. Marco contextual

Para entender, estructurar y desgranar cada paso que se ha dado hasta obtener los resultados que posteriormente se mostrarán es necesario entender básicamente el funcionamiento de este centro educativo. El hecho de ser de modalidad *online* hace que el trabajo diario de todas las partes implicadas (alumnos, profesores, directivos y administrativos) difiera en algunos aspectos con la enseñanza presencial.

Por tratarse de un centro de enseñanza de modalidad a distancia, la actividad diaria de cada usuario se lleva a cabo a través de una plataforma educativa virtual de formación, en este caso Moodle en su versión 2.5.9. Se trata de un sistema o conjunto de aplicaciones web diseñadas para la creación y seguimiento de cursos virtuales de formación, permitiendo, por parte del alumno, la entrega de tareas y consulta de contenidos online y, por parte del provesor, el seguimiento o gestión de la actividad diaria de sus alumnos.

Aunque la **puerta virtual de acceso** al centro es la misma para cada parte implicada, cada una tiene un *rol* específico según sus funciones. Así el primer paso para acceder es identificarse en la página principal la cuál se muestra en la siguiente figura:

Figura 1.1: Acceso

Una vez dentro, cada **aula** contiene una estructura más o menos similar organizada para facilitar al alumno la realización de un curso. En la siguiente figura se muestra un fragmento del contenido de la misma, en ella vemos desde una sección de vídeos para orientarlo, gestión de mensajes de correo electrónico, foros de discusión sobre distintos temas, etc.

Figura 1.2: Fragmento de un aula

Tanto el profesor como la directiva disponen de un **Menú de Administración** con varias opciones, la que nos interesa aparece como **Informe tutores** de esta forma:

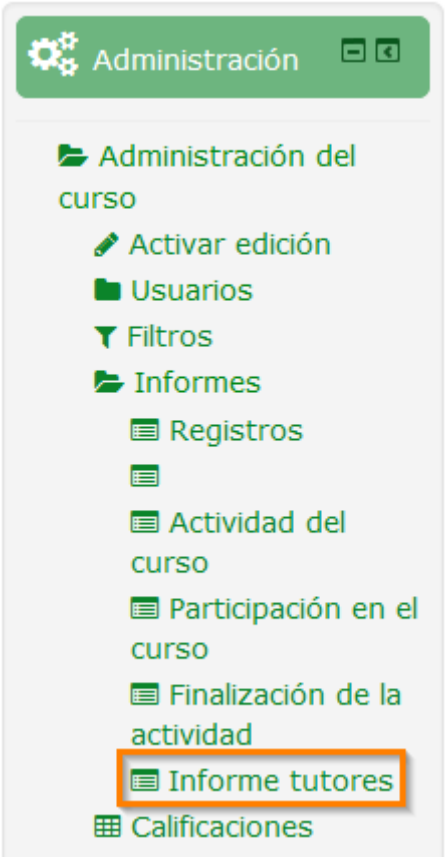


Figura 1.3: Menú de Administración

Haciendo clic en dicha opción, el sistema nos muestra un **listado ordenado de todos los alumnos** de ese aula y seguidamente haciendo clic en cada uno de ellos accedemos a su ficha (expediente) personal en la que el profesor dispone de información según su actividad desde el principio de curso. A modo de ejemplo, se muestra un fragmento de la ficha personal de un alumno concreto:

Ficha resumen del alumnado

FOTO

Nombre: [redacted]

Apellidos: [redacted]

DNI: [redacted] **Teléfono:** [redacted]

Primer acceso: jueves, 18 de septiembre de 2014, 21:22 (1 año 250 días)

Último acceso: domingo, 22 de mayo de 2016, 19:24 (3 días 21 horas)

Módulo/Materia	TRI_1	TRI_2	TRI_3	Correos	Foros	Cuestionarios on-line	Nota final Jun/Sept	1Med/1Sen	2Med/2Sen	3Med/3Sen
2º Bach - Historia de la Filosofía [redacted]	TT 4 TE 3 TC 3 TA 3 NEP Apto NRJ SC NRS SC	TT 4 TE 3 TC 4 TA 3 NEP Apto NRJ SC NRS SC	TT 5 TE 4 TC 6 TA 4 NEP SC NRJ SC NRS SC	CE 0 CR 0	FPE 4 FDC 0 FPL 19	EMR 0 EMA 0	SC/SC	6.2/6	6.53/7	5.28/0
2º Bach - Lengua Castellana y Literatura [redacted]	TT 5 TE 5 TC 13 TA 3 NEP Apto	TT 5 TE 5 TC 7 TA 4 NEP Apto	TT 5 TE 4 TC 4 TA 4 NEP SC NRJ SC	CE 0 CR 1	FPE 3 FDC 1 FPL 41	EMR 0 EMA 0	SC/SC	5.11/5	5.15/5	5.3/0

Figura 1.4: Ficha resumen del alumnado

Para aclarar las siglas de la figura anterior, al final de cada ficha encontramos la **leyenda** de la siguiente forma:

Figura 1.5: Leyenda de Ficha resumen del alumnado

Las siglas que aparecen hacen referencia a las distintas variables que se obtienen fruto de la actividad diaria del alumnado en cada alula. Más adelante se detallarán cada una de ellas en cuanto a sus tipos concretos y los distintos valores que pueden tomar.

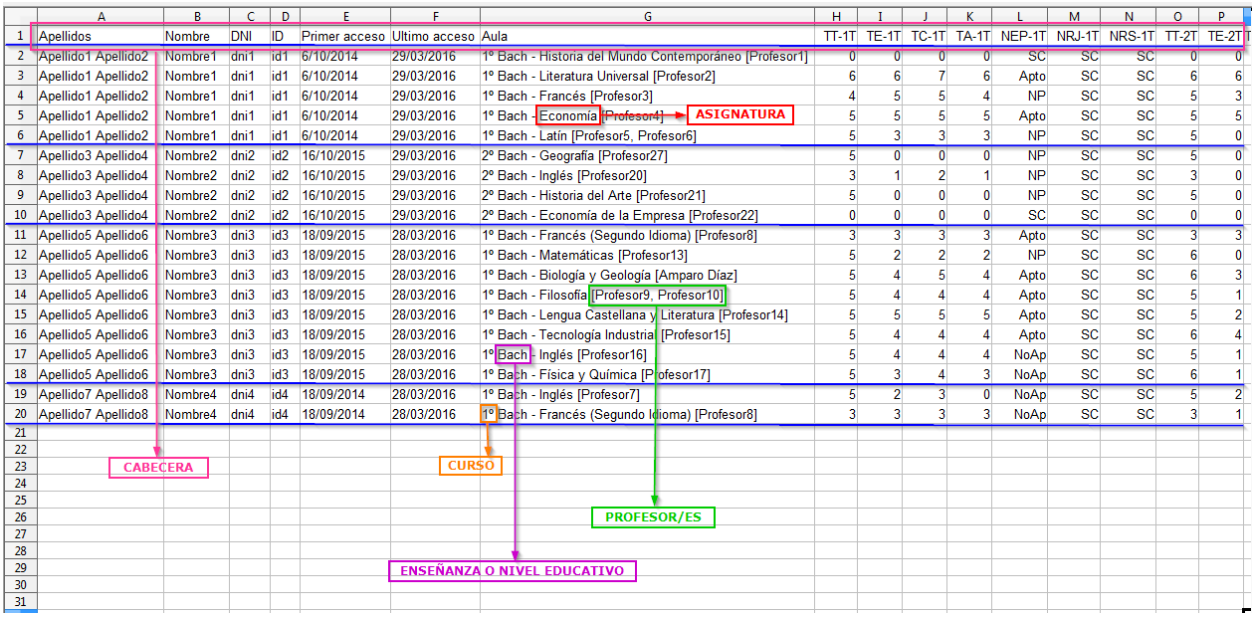
Aunque la parte que a continuación se explica excede de las herramientas y técnicas estadísticas, es necesario resaltar su importancia, puesto que ha sido necesario construir lo que llamamos un **sistema unificado de descarga** de cada ficha personal de cada alumno en cada aula y a su vez de cada curso y enseñanza. Sin entrar en detalles, se han elaborado varios **scripts de linux** (shell scripts) para, a partir de los datos de acceso de cada profesor, acceder **automáticamente** a sus aulas respectivas, descargar cada ficha personal de cada uno de sus alumnos, y **unificar** en un varios archivos CSV los datos de todas las fichas organizando por enseñanza y curso.

A modo de resumen, el procedimiento ha consistido en obtener, a partir del código fuente en formato HTML, de cada ficha personal los datos útiles que visualmente aparecen en la ficha del alumno, realizando búsquedas de patrones en dicho código para ir construyendo los ficheros CSV anteriormente citados, detalle que podemos ver en la figura que se muestra a continuación:


```
<tr>
<th class="header c0" style="text-align:left;width:20%;" scope="col"><a href="http://
<th class="header c1" style="text-align:center;width:15%;" scope="col">DNI</th>
<th class="header c2" style="text-align:center;width:5%;" scope="col">Cursos</th>
<th class="header c3" style="text-align:center;width:15%;" scope="col">Tareas</th>
<th class="header c4" style="text-align:center;width:15%;" scope="col">Correos</th>
<th class="header c5" style="text-align:center;width:15%;" scope="col">Foros</th>
<th class="header c6" style="text-align:center;width:15%;" scope="col">Cuestionarios<br> on-line</th>
<th class="header c7 lastcol" style="text-align:center;width:15%;" scope="col">Nota exámenes presenciales</th>
</tr>
</thead>
<tbody><tr class="r0">
<td class="cell c0" style="text-align:left;width:20%;"><a href="http://
<td class="cell c1" style="text-align:center;width:15%;">26047668f</td>
<td class="cell c2" style="text-align:center;width:5%;">9</td>
<td class="cell c3" style="text-align:center;width:15%;">TE <b>34</b><br/>TC <b>35</b><br/>TA <b>33</b></td>
<td class="cell c4" style="text-align:center;width:15%;">CE <b>9</b><br/>CR <b>29</b></td>
<td class="cell c5" style="text-align:center;width:15%;">FPE <b>6</b><br/>FDC <b>3</b><br/>FPL <b>70</b></td>
<td class="cell c6" style="text-align:center;width:15%;">EMR <b>6</b><br/>EMA <b>4</b></td>
<td class="cell c7 lastcol" style="text-align:center;width:15%;">EPR <b>13</b><br/>EPA <b>7</b></td>
</tr>
<tr class="r1">
<td class="cell c0" style="text-align:left;width:20%;"><a href="http://
<td class="cell c1" style="text-align:center;width:15%;">77934433z</td>
<td class="cell c2" style="text-align:center;width:5%;">10</td>
<td class="cell c3" style="text-align:center;width:15%;">TE <b>28</b><br/>TC <b>34</b><br/>TA <b>24</b></td>
<td class="cell c4" style="text-align:center;width:15%;">CE <b>0</b><br/>CR <b>16</b></td>
<td class="cell c5" style="text-align:center;width:15%;">FPE <b>0</b><br/>FDC <b>0</b><br/>FPL <b>37</b></td>
<td class="cell c6" style="text-align:center;width:15%;">EMR <b>1</b><br/>EMA <b>1</b></td>
<td class="cell c7 lastcol" style="text-align:center;width:15%;">EPR <b>6</b><br/>EPA <b>1</b></td>
</tr>
<tr class="r0">
```

Figura 1.6: Código fuente de la ficha resumen del alumnado

Tras realizar el proceso que implica el sistema unificado de descarga obtenemos, como ya se ha comentado, los archivos CSV anteriormente mencionados. En la siguiente figura podemos ver un fragmento de uno de ellos:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Apellidos	Nombre	DNI	ID	Primer acceso	Ultimo acceso	Aula	TT-1T	TE-1T	TC-1T	TA-1T	NEP-1T	NRJ-1T	NRS-1T	TT-2T	TE-2T
2	Apellido1 Apellido2	Nombre1	dni1	id1	6/10/2014	29/03/2016	1º Bach - Historia del Mundo Contemporáneo [Profesor1]	0	0	0	0	SC	SC	SC	0	0
3	Apellido1 Apellido2	Nombre1	dni1	id1	6/10/2014	29/03/2016	1º Bach - Literatura Universal [Profesor2]	6	6	7	6	Apto	SC	SC	6	6
4	Apellido1 Apellido2	Nombre1	dni1	id1	6/10/2014	29/03/2016	1º Bach - Francés [Profesor3]	4	5	5	4	NP	SC	SC	5	3
5	Apellido1 Apellido2	Nombre1	dni1	id1	6/10/2014	29/03/2016	1º Bach - Economía [Profesor4]	5	5	5	5	Apto	SC	SC	5	5
6	Apellido1 Apellido2	Nombre1	dni1	id1	6/10/2014	29/03/2016	1º Bach - Latín [Profesor5, Profesor6]	5	3	3	3	NP	SC	SC	5	0
7	Apellido3 Apellido4	Nombre2	dni2	id2	16/10/2015	29/03/2016	2º Bach - Geografía [Profesor27]	5	0	0	0	NP	SC	SC	5	0
8	Apellido3 Apellido4	Nombre2	dni2	id2	16/10/2015	29/03/2016	2º Bach - Inglés [Profesor20]	3	1	2	1	NP	SC	SC	3	0
9	Apellido3 Apellido4	Nombre2	dni2	id2	16/10/2015	29/03/2016	2º Bach - Historia del Arte [Profesor21]	5	0	0	0	NP	SC	SC	5	0
10	Apellido3 Apellido4	Nombre2	dni2	id2	16/10/2015	29/03/2016	2º Bach - Economía de la Empresa [Profesor22]	0	0	0	0	SC	SC	SC	0	0
11	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Francés (Segundo Idioma) [Profesor8]	3	3	3	3	Apto	SC	SC	3	3
12	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Matemáticas [Profesor13]	5	2	2	2	NP	SC	SC	6	0
13	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Biología y Geología [Amparo Díaz]	5	4	5	4	Apto	SC	SC	6	3
14	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Filosofía [Profesor9, Profesor10]	5	4	4	4	Apto	SC	SC	5	1
15	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Lengua Castellana y Literatura [Profesor14]	5	5	5	5	Apto	SC	SC	5	2
16	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Tecnología Industrial [Profesor15]	5	4	4	4	Apto	SC	SC	6	4
17	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Inglés [Profesor16]	5	4	4	4	NoAp	SC	SC	5	1
18	Apellido5 Apellido6	Nombre3	dni3	id3	18/09/2015	28/03/2016	1º Bach - Física y Química [Profesor17]	5	3	4	3	NoAp	SC	SC	6	1
19	Apellido7 Apellido8	Nombre4	dni4	id4	18/09/2014	28/03/2016	1º Bach - Inglés [Profesor7]	5	2	3	0	NoAp	SC	SC	5	2
20	Apellido7 Apellido8	Nombre4	dni4	id4	18/09/2014	28/03/2016	1º Bach - Francés (Segundo Idioma) [Profesor8]	3	3	3	3	NoAp	SC	SC	3	1

Figura 1.7: Fragmento Informe de tutores

En este sentido, remarcar que se ha intentado extraer la máxima información posible de cara a una mayor riqueza en los datos fuente de los que partiremos para su tratamiento. Aunque hay muchas variables que serían muy interesantes (tiempo de permanencia en la plataforma, enlaces más consultados, hora habitual de acceso, etc) y que en este momento no están disponibles, en el futuro se pretende seguir trabajando para ir adquiriéndolas.

A modo de resumen, la situación concreta sobre la que se ha realizado todo el trabajo es la siguiente:

- **Población**: datos originados fruto de la actividad diaria del alumnado de un centro educativo.
- **Tipo de centro educativo**: público de enseñanza a distancia (*online*).
- **Tipo de alumnado**: Adultos mayores edad, salvo excepciones justificadas a partir de 16 años.
- **Tipo de profesorado**: De cada especialidad de enseñanza secundaria y con buenos conocimientos de las TIC.
- **Plataforma de gestión y seguimiento académico**: Moodle (versión 2.5.9).
- **Enseñanzas**: 1º y 2º de Bachillerato, nivel 1 y 2 (cursos 1º y 2º) de ESA (ESO para adultos).

1.2. Objetivos propuestos

En un primer momento, el único objetivo que tenía el centro era el de resumir pequeñas y diversas fuentes de datos que cada unidad funcional del centro (departamentos, profesores, administrativos y alumnos) aportaba por su cuenta. Esta tarea, además de laboriosa, era fuente

de muchos errores lo cual hacía difícil el tomar algunas decisiones y provocaba incertidumbre, restando credibilidad a los resultados que se obtenían.

Los objetivos propuestos a llevar a cabo han sido principalmente dos:

1. **Explotación descriptiva de datos** para conocer la realidad de la actividad diaria *(qué está pasando)*.
2. **Construir modelos de predicción** para tomar decisiones con antelación *(qué es probable que pase)*.

Aunque el primer gran avance ha sido la obtención de unos primeros datos para su explotación estadística, el siguiente paso es intentar ampliar el número de variables que aporten más datos del seguimiento del alumnado (tiempo de permanencia en la plataforma desde que entra hasta que sale, enlaces consultados, días de la semana y horario en el que normalmente accede, etc) con el fin de elaborar informes sobre otras características de interés, tanto para profesorado como para el equipo directivo e incluso el alumnado(Andrés Muñoz Ortega, s.f.).

Capítulo 2

Explotación descriptiva de datos

Tal y como se expone en (Calot 1988), el fin de la Estadística Descriptiva es describir con los medios apropiados, separar lo esencial, resumirlo y medirlo. Conceptos de esta parte de la Estadística como los distintos tipos de variables que podemos medir, las tablas estadísticas para organizar los datos, o las representaciones gráficas son básicos y de suma importancia en todo trabajo estadístico.

En este capítulo se explica detalladamente el proceso referente a la explotación descriptiva de datos, para ello debemos partir de la estructura general en la que se encuentran organizados. Dicha estructura, difiere en algunos matices según la enseñanza (ESA o Bachillerato) y curso (1º o 2º). Sin embargo, el 90 % es común y por tanto, será esa la línea estructural que se seguirá tal y como se detallará en el apartado siguiente.

Una vez clara la estructura organizativa, el siguiente paso es la preparación previa de los datos fuente para poder ser tratados computacionalmente, algunos problemas de codificación de ficheros son expuestos en este capítulo, concretamente en el apartado relativo a la *Preparación de datos fuente*.

Tras finalizar el paso anterior, se cuenta con lo que llamamos **datos disponibles**, los cuales se especificarán para saber con qué posibles variables contamos para cada caso concreto (alumno). Será necesario comprobar, y en caso necesario, establecer los tipos adecuados de todas las variables (numérico, factor, carácter, fecha, etc).

Seguidamente es preciso realizar un primer análisis estadístico descriptivo básico para determinar en qué rango de valores se mueve cada variable, así como algunas características de resumen (media, desviación estándar, etc) y varias representaciones gráficas que ayuden a tener un mejor concepto general de la realidad de la actividad del centro.

Por último, se muestra una de las salidas (en formato PDF) impresas del **informe automático de estadísticas descriptivas** (para 1º de Bachillerato). El valor de dicho informe para el centro es bastante importante y supone un antes y un después a la hora de obtener información sobre la actividad diaria del alumnado durante el curso.

2.1. Estructura general de los datos

Los datos con los que contamos están organizados según la siguiente estructura general de **bloques**:

Enseñanza –> Curso –> Aula –> Tareas –> Calificaciones

A su vez, cada uno de estos cinco bloques se compone de varias **modalidades** de la siguiente forma:

- Enseñanza:
  - Bachillerato
  - ESA
- Curso:
  - Primero
  - Segundo
- Aula:
  - Aula1[Profesor/es 1/varios]
  - Aula2[Profesor/es 2/varios]
  - ...
  - Aulak[Profesor/es m/varios]
- Tareas:
  - Tarea individual 1
  - Tarea individual 2
  - ...
  - Tarea individual p (p = 3xt, siendo t el nº de trimestres)
  - Tarea colaborativa 1
  - Tarea colaborativa 2
  - ...
  - Tarea colaborativa q (p = 1xt, siendo t el nº de trimestres)
  - Tarea global 1
  - Tarea global 2
  - ...
  - Tarea global r (p = 1xt, siendo t el nº de trimestres)
- Calificaciones (individuales: 50 %, colaborativas: 20 %, global: 30 %, presenciales: apto/no apto):
  - Calificación tarea individual 1
  - Calificación tarea individual 2
  - ...
  - Calificación tarea individual p
  - Calificación tarea colaborativa 1

- Calificación tarea colaborativa 2
- ...
- Calificación tarea colaborativa q
- Calificación tarea global 1
- Calificación tarea global 2
- ...
- Calificación tarea global r
- Media tareas individuales
- Media tareas colaborativas
- Media tareas globales
- Calificación prueba presencial 1
- Calificación prueba presencial 2
- Calificación prueba presencial 3

Un *aula* se identifica por un curso, enseñanza, asignatura y profesor. Las que comienzan por *Nivel* hacen referencia a la ESA.

Las *tareas* para cada trimestre se dividen en individuales, colaborativas, globales y presenciales, así para el caso de segundo de bachillerato se desglosan para cada aula en:

- Tareas 1º Trimestre: 3 tareas individuales, 1 tarea colaborativa, 1 tarea global, 1 tarea presencial.
- Tareas 2º Trimestre: 3 tareas individuales, 1 tarea colaborativa, 1 tarea global, 1 tarea presencial.
- Tareas 3º Trimestre: 3 tareas individuales, 1 tarea colaborativa, 1 tarea global, 1 tarea presencial.
- Tarea presencial de junio.
- Tarea presencial de septiembre.

Los pesos concretos de cada *tipo de tarea* (individual, colaborativa, global y presencial) varía en función del curso y la enseñanza y no se detallarán por ser irrelevantes para nuestro objetivo.

Los apartados referentes al grupo de *Calificaciones* se obtendrán ponderando según los pesos mencionados para cada tipo de tarea.

A modo de ejemplo, se muestra un fragmento de la estructura general descrita:

- Enseñanza: Bachillerato
  - Curso: 1º
  - Aula:
    - 1º Bach - Biología y Geología [Profesor 1]
    - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
    - 1º Bach - Cultura Audiovisual [Profesor 2]
    - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.

        - ...
    - Curso: 2º
    - Aula:
        - 2º Bach - Geografía [Profesor 10]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - 2º Bach - Geografía [Profesor 11]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - 2º Bach - Literatura Universal [Profesor 12]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - ...

- Enseñanza: ESA
    - Curso: 1º
    - Aula:
        - Nivel I - Ámbito científico-tecnológico [Profesor 30]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - Nivel I - Ámbito científico-tecnológico [Profesor 31]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - ...

- Enseñanza: ESA
    - Curso: 2º
    - Aula:
        - Nivel II - Ámbito científico-tecnológico [Profesor 40]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - Nivel II - Ámbito social [Profesor 41]
        - Tareas: Tareas 1º Tri, Tareas 2º Tri, Tareas 3º Tri.
        - ...

2.2. Preparación de datos fuente

La propia plataforma Moodle es la encargada de recopilar la gran cantidad de datos que se generan fruto de las actividades que realizan los distintos usuarios. Entre otras características, el rol del usuario determina qué puede gestionar, y por tanto, los datos resultantes aparecerán filtrados según esta función.

Con el sistema unificado de descarga se consiguen una serie de archivos .csv nombrados como **InformeTutoresCursoEnseñanza.csv** diferenciados por enseñanza y curso, concretamente los siguientes:

1. InformeTutores1BAC.csv : Para primero de bachillerato.
2. InformeTutores2BAC.csv : Para segundo de bachillerato.
3. InformeTutores1ESA.csv : Para primero de ESA.

4. InformeTutores2ESA.csv : Para segundo de ESA.

Los informes 1 y 3 contienen los datos de alumnos matriculados solamente en primero. Los informes 2 y 4 hacen referencia al segundo curso o nivel, en ellos aparecen datos de alumnos que están matriculados como último curso en segundo e incluye alumnos con materias pendientes de primero.

En principio parecía que sólo con tener los archivos fuente podíamos ponernos "manos a la obra", sin embargo, una parte importante del trabajo se ha centrado en preparar dichos ficheros para tratarlos posteriormente.

Un primer obstáculo ha sido la **codificación**, tanto la que por defecto traían dichos archivos (ISO-8859-1) como la conversión al tipo adecuado para su importación y tratamiento (UTF-8). Seguidamente, para tener diferenciados los alumnos por curso, ha sido necesario extraer de los informes de segundo curso (informes 2 y 4) los alumnos de primero y añadirlos a sus correspondientes informes de primer curso (informes 1 y 3). Centrándonos en un curso y enseñanza concreta, el paso siguiente ha sido **eliminar casos no válidos**, estos son, referentes a las llamadas *Aulas Piloto* o experimentales, asignaturas cuatrimestrales (por no seguir la línea temporal de trimestres fijada), con valores faltantes, etc.

Una vez filtrados los casos que se consideran válidos, ha sido necesario renombrar la mayoría de las variables puesto que al importar desde R algunas tenían signos de puntuación y/o caracteres especiales que impedían hacer referencia a ellas de una forma lógica. Para este tipo de tareas se ha usado el paquete **rattle** de R cuyo entorno gráfico facilita enormemente la depuración previa de datos(Williams 2011).

No podemos olvidar el comprobar el **tipo de cada variable** (numérico, factor, carácter, lógico, fecha) y convertir según el caso algunas de ellas al tipo correcto. Aquí apuntamos el tipo *fecha* como necesario para poder construir nuevas variables a partir de las originales que permitan realizar cálculos numéricos (días pasados desde una fecha determinada, etc).

A continuación se muestra un fragmento (cinco primeros casos o filas) de la **matriz de datos** de uno de los cuatro archivos .csv que obtenemos del sistema unificado de descarga y sobre la que vamos a trabajar. Se trata de un archivo compuesto por una serie de variables relativas al seguimiento de cada alumno:

Tabla 2.1: Informe de Tutores 1º Bachillerato

	Apellidos	Nombre	DNI	ID
1	Apellido1 Apellido2	Nombre 1	dni1	id1
2	Apellido1 Apellido2	Nombre 1	dni1	id1
3	Apellido3 Apellido4	Nombre 2	dni2	id2
4	Apellido3 Apellido4	Nombre 2	dni2	id2
5	Apellido3 Apellido4	Nombre 2	dni2	id2

	Primer.acceso	Ultimo.acceso	Aula
1	21/10/2015	8/03/2016	1º Bach - Filosofía [Profesor 7, Profesor 8]
2	21/10/2015	8/03/2016	1º Bach - Francés (Segundo Idioma) [Profesor 10]
3	18/09/2015	28/03/2016	1º Bach - Biología y Geología [Profesor 1]
4	18/09/2015	28/03/2016	1º Bach - Dibujo Técnico [Profesor 3]
5	18/09/2015	28/03/2016	1º Bach - Inglés [Profesor 18]

	TT.1T	TE.1T	TC.1T	TA.1T	NEP.1T	NRJ.1T	NRS.1T
1	5	0	0	0	NP	SC	SC
2	3	0	0	0	NP	SC	SC
3	5	5	9	3	Apto	SC	SC
4	6	6	6	2	NoAp	SC	SC
5	5	4	4	4	NoAp	SC	SC

	TT.2T	TE.2T	TC.2T	TA.2T	NEP.2T	NRJ.2T	NRS.2T
1	5	0	0	0	NP	SC	SC
2	3	0	0	0	NP	SC	SC
3	6	6	6	3	Apto	SC	SC
4	6	5	5	5	NoAp	SC	SC
5	5	4	4	3	NoAp	SC	SC

	TT.3T	TE.3T	TC.3T	TA.3T	NEP.3T	NRJ.3T	NRS.3T
1	5	0	0	0	SC	SC	SC
2	3	0	0	0	SC	SC	SC
3	6	1	1	1	SC	SC	SC
4	6	0	0	0	SC	SC	SC
5	5	0	0	0	SC	SC	SC

	CE	CR	FPE	FDC	FPL	EMR	EMA
1	0	0	0	0	4	0	0
2	0	4	0	0	0	0	0
3	2	3	1	1	45	0	0
4	3	8	0	0	11	0	0
5	3	5	0	0	7	4	2

	FINAL.JUNIO	FINAL.SEPTIEMBRE
1	SC	SC

	FINAL.JUNIO	FINAL.SEPTIEMBRE
2	SC	SC
3	SC	SC
4	SC	SC
5	SC	SC

	X1T	X1T.SENECA	X2T	X2T.SENECA	X3T	X3T.SENECA
1	0.00	0	0.00	0	0.00	0
2	0.00	0	0.00	0	0.00	0
3	5.79	6	5.37	5	0.99	0
4	5.20	4	5.03	4	0.00	0
5	5.84	4	4.29	4	0.00	0

2.3. Datos disponibles

El total aproximado de casos válidos con los que contamos para trabajar son:

- 1300 casos relativos a alumnos de 1º de Bachillerato.
- 1800 casos relativos a alumnos de 2º de Bachillerato.
- 200 casos relativos a alumnos de nivel 1 de ESA.
- 500 casos relativos a alumnos de nivel 2 de ESA.

El total de variables y tipo se resumen en:

- de tipo carácter:
  - Apellidos: Apellidos del alumno.
  - Nombre: Nombre del alumno.
- de tipo factor:
  - DNI: Documento Nacional de Identidad del alumno.
  - ID: Identificador único de alumno en la plataforma Moodle.
  - Aula: Campo múltiple que contiene implícitas las variables: Curso (factor), Enseñanza (factor), Asignatura (factor), Profesor/es (factor).
  - NEP.1T: Nota (calificación) obtenida por el alumno en el primer trimestre.

- NRJ.1T: Nota (calificación) de recuperación en junio obtenida por el alumno del primer trimestre.
- NRS.1T: Nota (calificación) de recuperación en septiembre obtenida por el alumno del primer trimestre.
- NEP.2T: Nota (calificación) obtenida por el alumno en el primer trimestre.
- NRJ.2T: Nota (calificación) de recuperación en junio obtenida por el alumno del primer trimestre.
- NRS.2T: Nota (calificación) de recuperación en septiembre obtenida por el alumno del primer trimestre.
- NEP.3T: Nota (calificación) obtenida por el alumno en el primer trimestre.
- NRJ.3T: Nota (calificación) de recuperación en junio obtenida por el alumno del primer trimestre.
- NRS.3T: Nota (calificación) de recuperación en septiembre obtenida por el alumno del primer trimestre.

- de tipo cuantitativo discreto:

  - TT.1T: Nº máximo de tareas que el alumno debe entregar en el primer trimestre.
  - TE.1T: Nº de tareas entregadas por el alumno en el primer trimestre.
  - TC.1T: Nº de tareas corregidas al alumno por su profesor en el primer trimestre.
  - TA.1T: Nº de tareas aprobadas por el alumno en el primer trimestre.
  - TT.2T: Nº máximo de tareas que el alumno debe entregar en el primer trimestre.
  - TE.2T: Nº de tareas entregadas por el alumno en el primer trimestre.
  - TC.2T: Nº de tareas corregidas al alumno por su profesor en el primer trimestre.
  - TA.2T: Nº de tareas aprobadas por el alumno en el primer trimestre.
  - TT.3T: Nº máximo de tareas que el alumno debe entregar en el primer trimestre.
  - TE.3T: Nº de tareas entregadas por el alumno en el primer trimestre.
  - TC.3T: Nº de tareas corregidas al alumno por su profesor en el primer trimestre.
  - TA.3T: Nº de tareas aprobadas por el alumno en el primer trimestre.
  - CE: Correos enviados por el alumno.
  - CR: Correos recibidos por el alumno.
  - FPE: Mensajes (posts) enviados por el alumno a los foros.
  - FDC: Mensajes (posts) creados por el alumno en los foros.
  - FPL: Mensajes (posts) leídos por el alumno en los foros.
  - EMR: Exámenes presenciales realizados por el alumno.
  - EMA: Exámenes presenciales aprobados por el alumno.

  - FINAL.JUNIO: Nota final de junio.
  - FINAL.SEPTIEMBRE: Nota final de septiembre.

- de tipo cuantitativo continuo:

  - X1T: Nota media ponderada del alumno en el primer trimestre calculada automáticamente por la plataforma.
  - X1T.SENECA: Nota del alumno final en el primer trimestre revisada por el profesor.
  - X2T: Nota media ponderada del alumno en el segundo trimestre calculada automáticamente por la plataforma.
  - X2T.SENECA: Nota del alumno final en el segundo trimestre revisada por el profesor.
  - X3T: Nota media ponderada del alumno en el tercer trimestre calculada automáticamente por la plataforma.
  - X3T.SENECA: Nota del alumno final en el tercer trimestre revisada por el profesor.

- de tipo fecha:

  - Primer.acceso: Fecha en que un alumno accede a la plataforma por primera vez.
  - Ultimo.acceso: Fecha en que un alumno accede a la plataforma por última vez.

2.4. Análisis estadístico descriptivo inicial

Con el fin de resumir las principales características de las variables que vamos a usar, se procede a realizar un análisis descriptivo inicial o básico. Pretendemos obtener algunas medias de tendencia central y dispersión para hacernos una primera idea sobre el rango de valores en qué se mueve cada una de ellas. Para este apartado consideraremos los datos una vez terminada la segunda evaluación de primero de Bachillerato. El resumen es el siguiente:

```
##             mean         sd   IQR 0% 25% 50%   75% 100%    n
## CE     0.6738492  3.5463861  0.00  0   0 0.0  0.00   60 1021
## CR     2.8060725  3.8296575  4.00  0   0 1.0  4.00   42 1021
## FDC    0.4299706  0.9162171  0.00  0   0 0.0  0.00    6 1021
## FPE    1.5690500  3.5007537  2.00  0   0 0.0  2.00   38 1021
## FPL   20.1018609 41.3981316 21.00  0   0 5.0 21.00  549 1021
## TT.1T  4.9304603  0.7846959  0.00  3   5 5.0  5.00    6 1021
## TT.2T  5.1665034  0.8029934  1.00  3   5 5.0  6.00    6 1021
## TE.1T  2.0264447  2.0887856  4.00  0   0 1.0  4.00    6 1021
## TE.2T  1.3212537  2.0442720  3.00  0   0 0.0  3.00    6 1021
## TA.1T  1.8452498  2.0290898  4.00  0   0 1.0  4.00    6 1021
```

```
## TA.2T  1.1958864  1.9389766  2.00  0   0 0.0  2.00    6 1021
## TC.1T  2.2526934  2.4202587  4.00  0   0 2.0  4.00   14 1021
## TC.2T  1.4231146  2.2633513  3.00  0   0 0.0  3.00   12 1021
## X1T    3.0374927  3.3283862  6.29  0   0 1.5  6.29   10 1021
## X2T    1.9262488  3.0767488  3.90  0   0 0.0  3.90   10 1021
```

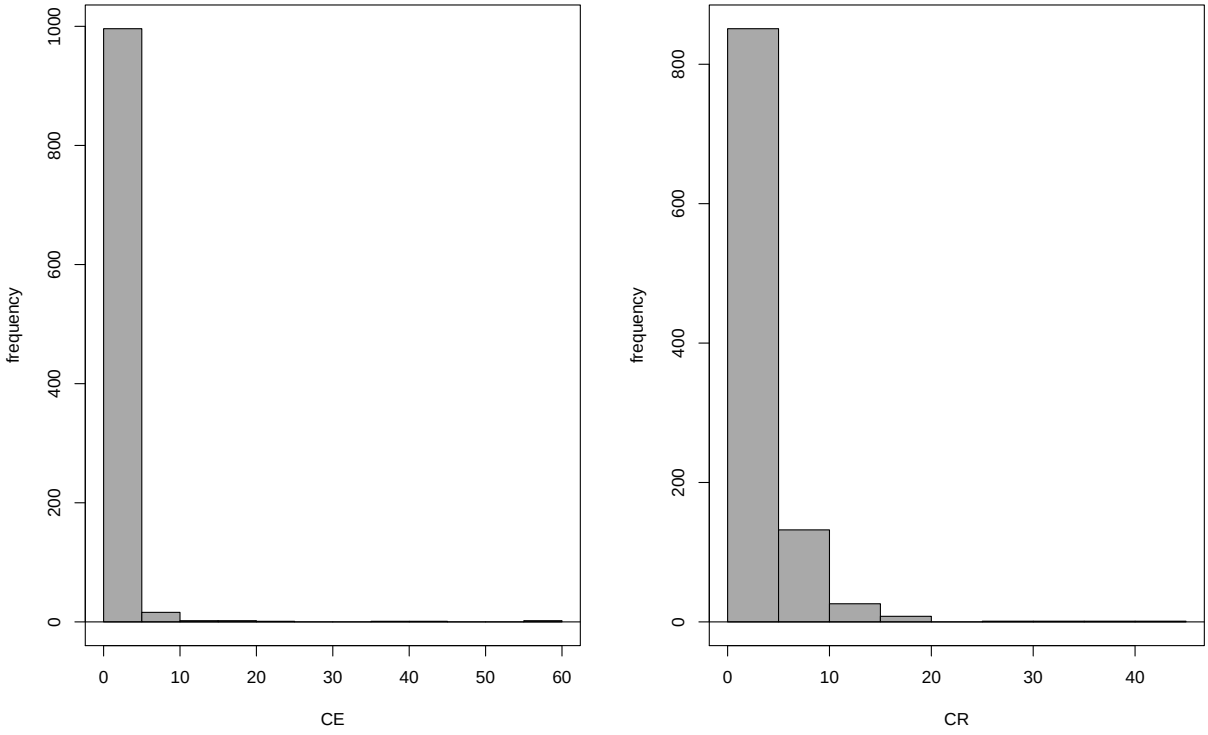


Figura 2.1: Histogramas Mensajes de correo

Observamos la poca actividad mostrada en estas dos variables.

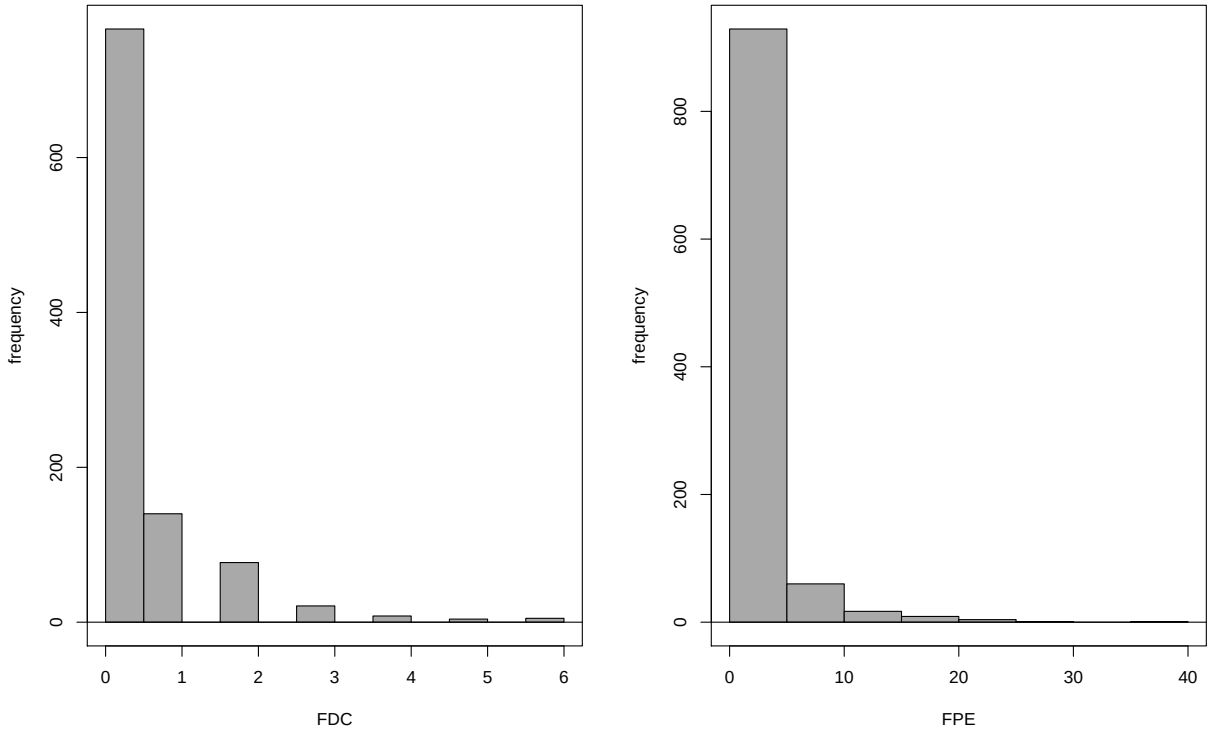


Figura 2.2: Histogramas Mensajes en foros

También las distribuciones de estas dos variables indican poca actividad, sobre todo la referente a la creación de nuevos temas de consulta.

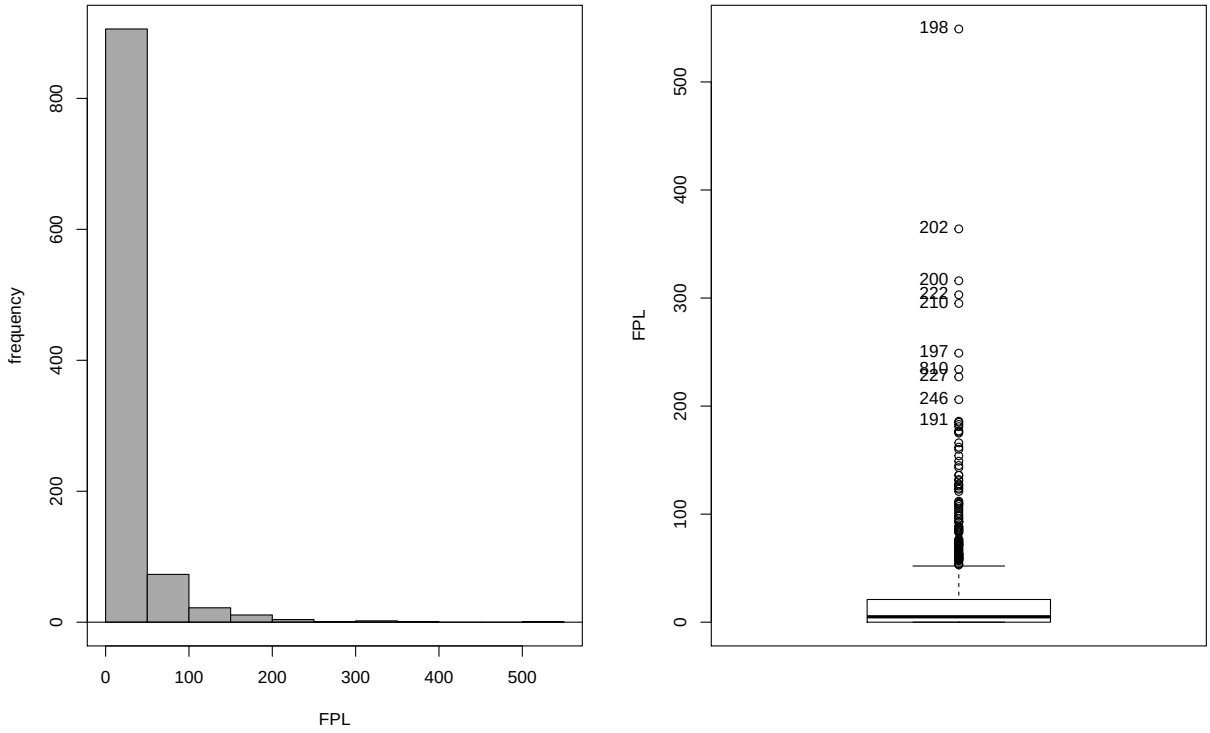


Figura 2.3: Histograma y Boxplot mensajes leídos en foros

```
##  [1] "198" "202" "200" "222" "210" "197" "810" "227" "246" "191"
```

Aunque la gráfica puede parecer similar a las anteriores, la marcada asimetría a la izquierda de la distribución de esta variable junto con su escala de valores es indicativa de que probablemente será influyente en la construcción de modelos que se verá más adelante. En el boxplot se representan además algunos *outliers*, los cuales se concretan justo debajo de la gráfica, que tendremos que analizar.

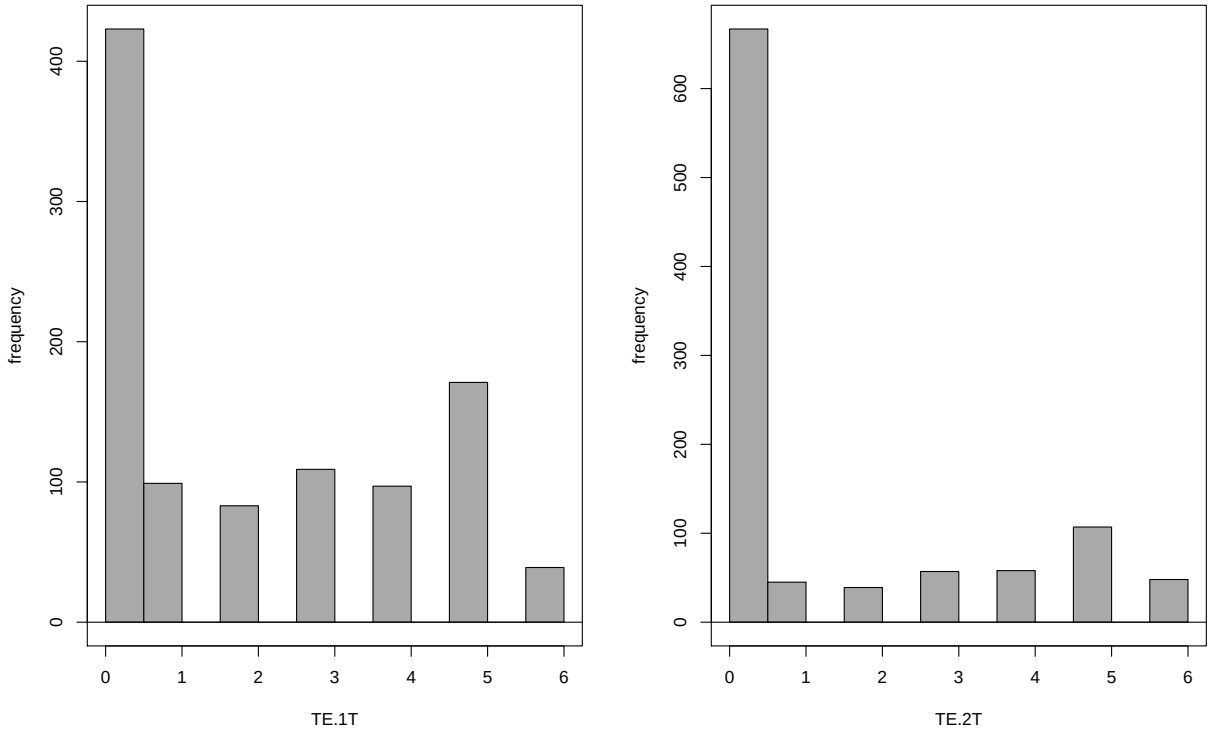


Figura 2.4: Histogramas Tareas entregadas

La distribución de las tareas entregadas en los dos trimestres es similar, aunque sí notamos como disminuye en la segunda evaluación con respecto a la primera.

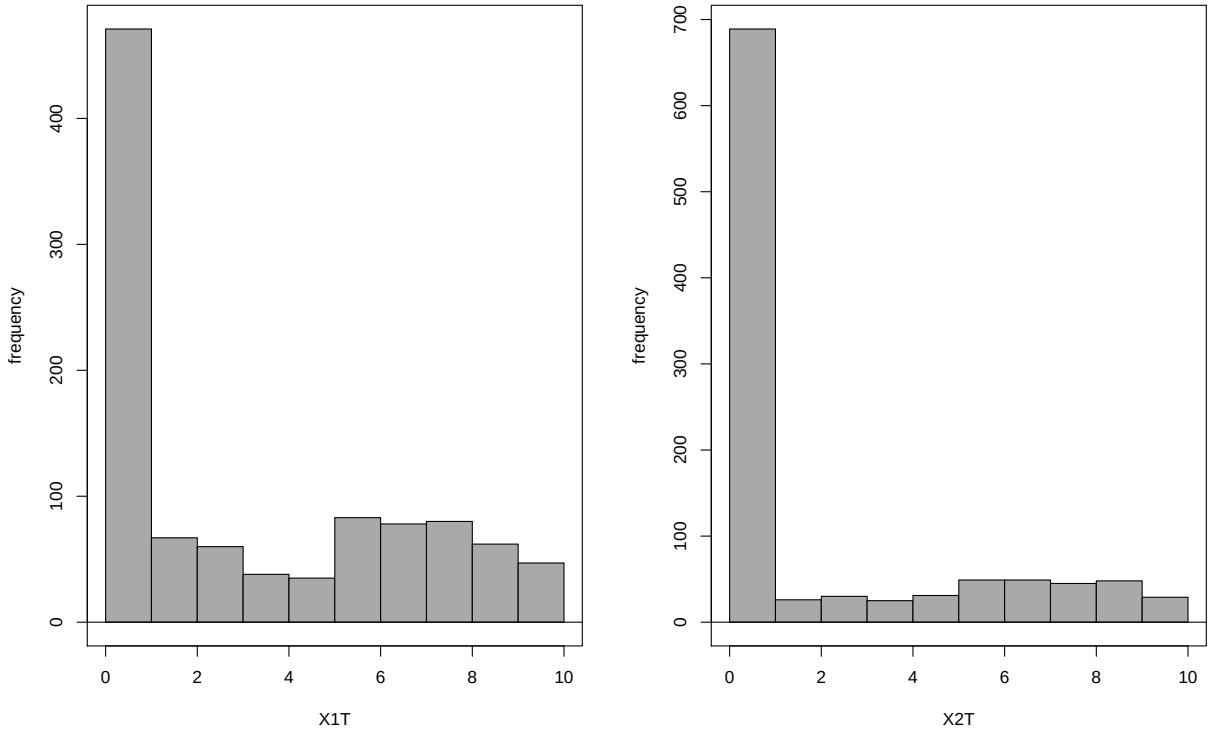


Figura 2.5: Histogramas Calificaciones

Lo más destacado es el alto grado de valores con calificación cero en ambos trimestres, indicativos de alumnos que o bien abandonan o son repetidores que no entregan tareas (y por tanto no se le califican) en este curso por tenerlas superadas en cursos anteriores.

En total contamos con aproximadamente (1300 + 1800 + 200 + 500) x 43 = 163.400 valores individuales, cifra considerable la cuál invita a estudiar la posible reducción de dimensonalidad de variables y casos.

Dejando a un lado las variables de tipo factor, si realizamos un Análisis de Componentes Principales los resultados obtenidos refuerzan la idea anterior(Cuadras 1991):

```
##
## Component loadings:
##                 Comp.1      Comp.2       Comp.3       Comp.4       Comp.5
## CE         -0.08952850  0.03333504  0.2668917933  0.590566294 -0.236889319
## CR         -0.08073077  0.07921319  0.3081268595  0.597017569 -0.115673230
## EMA        -0.08900937  0.40276860  0.4955782585 -0.279479915  0.013973589
## EMR        -0.08529769  0.40159090  0.5002566048 -0.280178317 -0.003149865
## FDC        -0.21210454 -0.14320048  0.0448342098 -0.087199027 -0.152600532
## FPE        -0.21951072 -0.05894251 -0.0887808216 -0.220458406 -0.302885092
## FPL        -0.19715760 -0.03545859 -0.0764361898 -0.195960577 -0.302053200
## TA.1T      -0.28908456 -0.02267695  0.0000338833  0.033273753  0.352272397
## TA.2T      -0.29588084 -0.04711347 -0.0272500842 -0.010719391 -0.181457975
## TC.1T      -0.27989499 -0.01696187  0.0035756914  0.020550765  0.341909675
## TC.2T      -0.29119027 -0.02713134 -0.0432486683 -0.026061201 -0.175400218
## TE.1T      -0.28960067 -0.02379176  0.0003075726  0.025065422  0.355829748
## TE.2T      -0.29794440 -0.03462823 -0.0277911821 -0.017811674 -0.172978603
## TT.1T      -0.01338486 -0.54700413  0.4014611231 -0.151620512  0.048294621
## TT.2T      -0.02158978 -0.57399283  0.3602769422 -0.054718032  0.077566941
## X1T        -0.28594280  0.07811022 -0.0912118207  0.082763962  0.323998910
## X1T.SENECA -0.28121036  0.05235444 -0.0799151472  0.086120538  0.249528915
## X2T        -0.29448045  0.02573834 -0.0796375839  0.008121448 -0.196702714
## X2T.SENECA -0.28832294 -0.01301917 -0.0825867463 -0.001384249 -0.220290479
##                 Comp.6       Comp.7       Comp.8      Comp.9       Comp.10
## CE         -0.12955887  0.219108548  0.6661053986 -0.05058910  0.011405493
## CR         -0.14999154 -0.145322934 -0.6781025955  0.10878694  0.061850325
## EMA         0.03220844  0.012935629  0.0134713637 -0.08099154  0.007151439
## EMR         0.03021928 -0.002601101  0.0300714981 -0.01004017 -0.034250986
## FDC        -0.33437106 -0.780183259  0.1730819415 -0.02498353 -0.365309508
## FPE        -0.46854924 -0.010099990  0.0117973277 -0.11150010  0.660501951
## FPL        -0.46294229  0.542771376 -0.1953681227  0.09049286 -0.477556778
## TA.1T      -0.09805336  0.019605563  0.0224312600  0.02496543  0.028845846
## TA.2T       0.29131676 -0.010754068 -0.0148804186  0.01332015 -0.002518807
## TC.1T      -0.10182898  0.048730235  0.1074213338  0.46871635  0.043775336
## TC.2T       0.27414995 -0.025522306  0.0142032505  0.32377873 -0.061643101
## TE.1T      -0.09822481  0.049598355  0.0330584396  0.20395395 -0.007821120
## TE.2T       0.27765705 -0.010181504 -0.0002994439  0.17253578 -0.045626779
## TT.1T       0.02655439  0.044047756 -0.0343677137  0.14128187  0.329619940
## TT.2T       0.10164772  0.126908768 -0.0384864868 -0.30762819 -0.265655476
```

```
## X1T         -0.07340536  0.042870832 -0.0117834081 -0.23070507  0.012336502
## X1T.SENECA -0.03302953  0.029485362 -0.0756760648 -0.57135900 -0.013725932
## X2T          0.27319575  0.007070584 -0.0506501951 -0.11114069  0.062365372
## X2T.SENECA  0.23282414  0.020417125 -0.0497728406 -0.23320003  0.069972281
##                Comp.11       Comp.12     Comp.13      Comp.14
## CE           0.04792330  7.168393e-05 -0.02939345  0.005210513
## CR          -0.05394155  1.036216e-03  0.01623956  0.001730679
## EMA          0.01227405 -9.360454e-02  0.03931012  0.696733266
## EMR         -0.05548293  9.526586e-02 -0.01667571 -0.696078636
## FDC          0.11163241 -3.631202e-02  0.02507304  0.009506448
## FPE         -0.34889895  7.521205e-02 -0.04274359  0.008261769
## FPL          0.19314562 -3.051068e-02 -0.02617647  0.001364844
## TA.1T        0.00809357 -3.508685e-01 -0.27661906 -0.025793282
## TA.2T       -0.01303932 -1.431548e-01 -0.17119851  0.019798827
## TC.1T       -0.11517055  2.394949e-01  0.51025825  0.027967264
## TC.2T       -0.12952746  3.784110e-01 -0.21490343  0.112091711
## TE.1T       -0.03869716 -1.230654e-01  0.03536336 -0.005798029
## TE.2T       -0.06897772  1.237443e-01 -0.22544085 -0.006441966
## TT.1T        0.59207339  1.209894e-03 -0.09978544 -0.029928135
## TT.2T       -0.57069255  1.621964e-02  0.09390994  0.016612313
## X1T         -0.05017031 -3.010513e-01 -0.22351474 -0.061653988
## X1T.SENECA  0.27034187  6.363249e-01 -0.02694030  0.043255122
## X2T          0.05238684 -1.851670e-01 -0.02984043 -0.050084346
## X2T.SENECA  0.17013980 -2.637236e-01  0.67757556 -0.077557759
##                  Comp.15      Comp.16      Comp.17       Comp.18
## CE          -0.0097853781 -0.003158809  0.010600594 -0.0103597717
## CR          -0.0048607525  0.004961333 -0.004777403  0.0102517953
## EMA          0.0122507967 -0.014111413  0.015544016 -0.0044453031
## EMR         -0.0200196870  0.002069596 -0.011930436  0.0108080881
## FDC          0.0432908666 -0.025595411  0.005633371 -0.0041084662
## FPE         -0.0234791997  0.037842033  0.019197806 -0.0007498672
## FPL          0.0316036205 -0.017246786 -0.025606455  0.0135169024
## TA.1T       -0.4362080606  0.138773158 -0.564791876 -0.2064036823
## TA.2T        0.2620608553  0.513895339 -0.154159943  0.5008793011
## TC.1T        0.3252297636 -0.055786545 -0.291840951  0.1329632886
## TC.2T       -0.3652599459 -0.439883771 -0.009500500 -0.0255519781
## TE.1T       -0.0706054389  0.324407787  0.664101760 -0.3022975219
## TE.2T       -0.0645057160  0.171817998  0.228484196  0.0440568133
## TT.1T        0.0432712572 -0.114683878  0.045620240  0.0558838390
## TT.2T        0.0103108842 -0.012609307 -0.020947540 -0.0487486894
## X1T          0.1844768398 -0.554508382  0.226483573  0.4323507954
## X1T.SENECA -0.0004742973  0.150980342 -0.048995534 -0.0424787859
## X2T          0.5377047946 -0.185705156 -0.129516875 -0.6255978434
## X2T.SENECA -0.4043263418 -0.067726494  0.064669525  0.0933145536
##                  Comp.19
```

```
## CE          0.0080058416
## CR         -0.0015232787
## EMA        -0.0380807888
## EMR         0.0443019507
## FDC        -0.0022064701
## FPE         0.0027946161
## FPL        -0.0002992213
## TA.1T      -0.0852648062
## TA.2T       0.3650892425
## TC.1T      -0.1209739180
## TC.2T       0.3874757065
## TE.1T       0.2699919164
## TE.2T      -0.7817973680
## TT.1T      -0.0066784982
## TT.2T       0.0002159451
## X1T        -0.0678702749
## X1T.SENECA  0.0139272543
## X2T         0.0428425014
## X2T.SENECA -0.0198096877
##
## Component variances:
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
## 9.64369509 1.91781027 1.86151280 1.56114380 1.18356911 1.04117849
##     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
## 0.54591728 0.29814431 0.24284133 0.17619956 0.16579911 0.09669214
##    Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18
## 0.08233337 0.06028994 0.03859331 0.03582643 0.02135423 0.01477101
##    Comp.19
## 0.01232845
##
## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     3.1054299 1.3848503 1.36437268 1.24945740
## Proportion of Variance 0.5075629 0.1009374 0.09797436 0.08216546
## Cumulative Proportion  0.5075629 0.6085003 0.70647464 0.78864010
##                            Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     1.08791962 1.02038154 0.73886215 0.54602592
## Proportion of Variance 0.06229311 0.05479887 0.02873249 0.01569181
## Cumulative Proportion  0.85093321 0.90573208 0.93446457 0.95015638
##                            Comp.9     Comp.10     Comp.11    Comp.12
## Standard deviation     0.49278933 0.419761313 0.407184366 0.31095359
## Proportion of Variance 0.01278112 0.009273661 0.008726269 0.00508906
## Cumulative Proportion  0.96293750 0.972211159 0.980937428 0.98602649
##                           Comp.13     Comp.14     Comp.15     Comp.16
## Standard deviation     0.286937917 0.245540100 0.196451795 0.189278705
```

```
## Proportion of Variance 0.004333335 0.003173155 0.002031227 0.001885601
## Cumulative Proportion  0.990359823 0.993532978 0.995564205 0.997449806
##                           Comp.17      Comp.18      Comp.19
## Standard deviation     0.146130875 0.1215360240 0.1110335321
## Proportion of Variance 0.001123907 0.0007774213 0.0006488655
## Cumulative Proportion  0.998573713 0.9993511345 1.0000000000
```

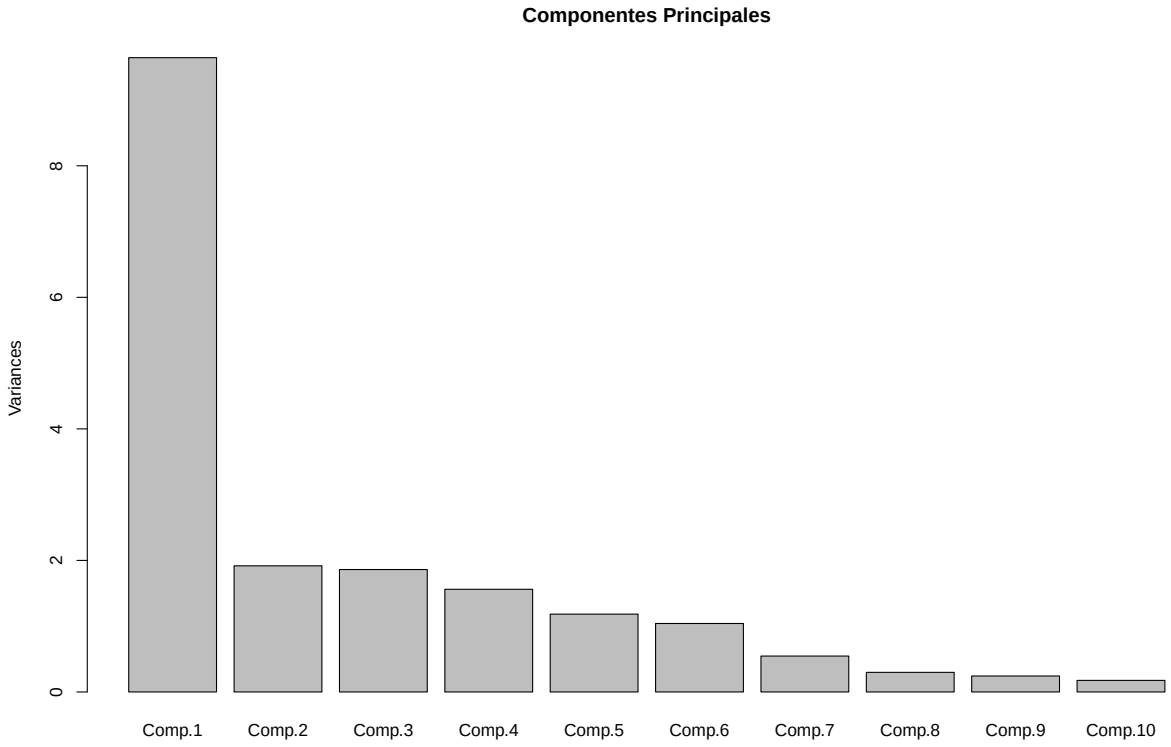


Figura 2.6: Componentes principales

La primera componente principal explica el 50 % de la varianza total, y con las tres primeras el 70 %, signo evidente de que existe información redundante y por tanto es factible plantearse una reducción de dimensionalidad considerando las componentes principales que deseemos en función de la precisión que queramos obtener en posibles resultados a partir de estas. Algunos resultados que podemos observar serían:

- La primera componente principal parece estar relacionada de una forma proporcionada con las tareas entregadas por el alumno y las calificaciones obtenidas.
- La segunda componente principal refleja proporcionalmente su correlación con los exámenes presenciales realizados y calificados.
- La cuarta parece centrarse más en las variables relativas a los correos enviados y recibidos.

- Etc.

Lo que sí parece claro en general es que cada componente principal *explica* un grupo de variables que en conjunto reflejarían una característica del alumno diferenciada de las demás (calificaciones, tareas, correos, mensajes en foros y exámenes presenciales).

La representación gráfica de cada característica diferenciada por trimestres es la siguiente:

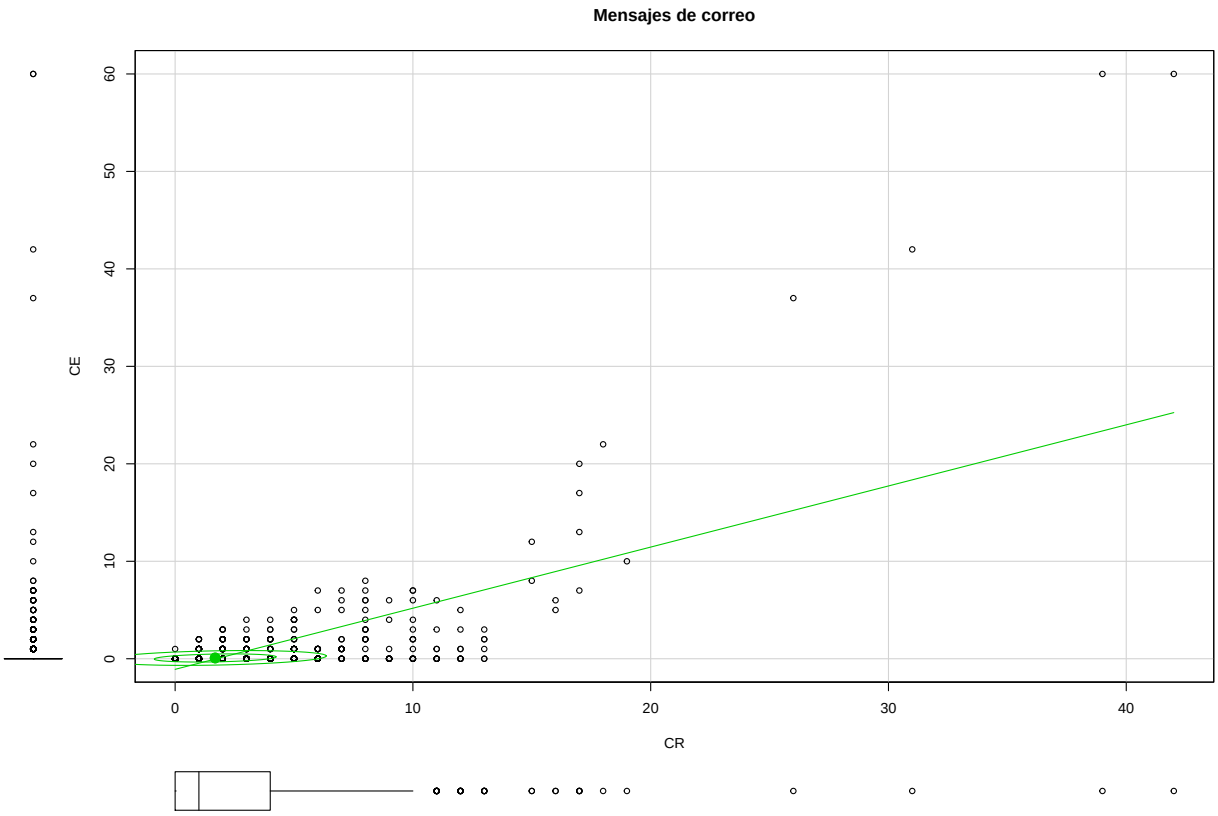


Figura 2.7: Mensajes de correo

Salvo algunas excepciones, comprobamos que la mayoría de los valores de los mensajes de correo (tanto CR, como CE) toman valores pequeños.

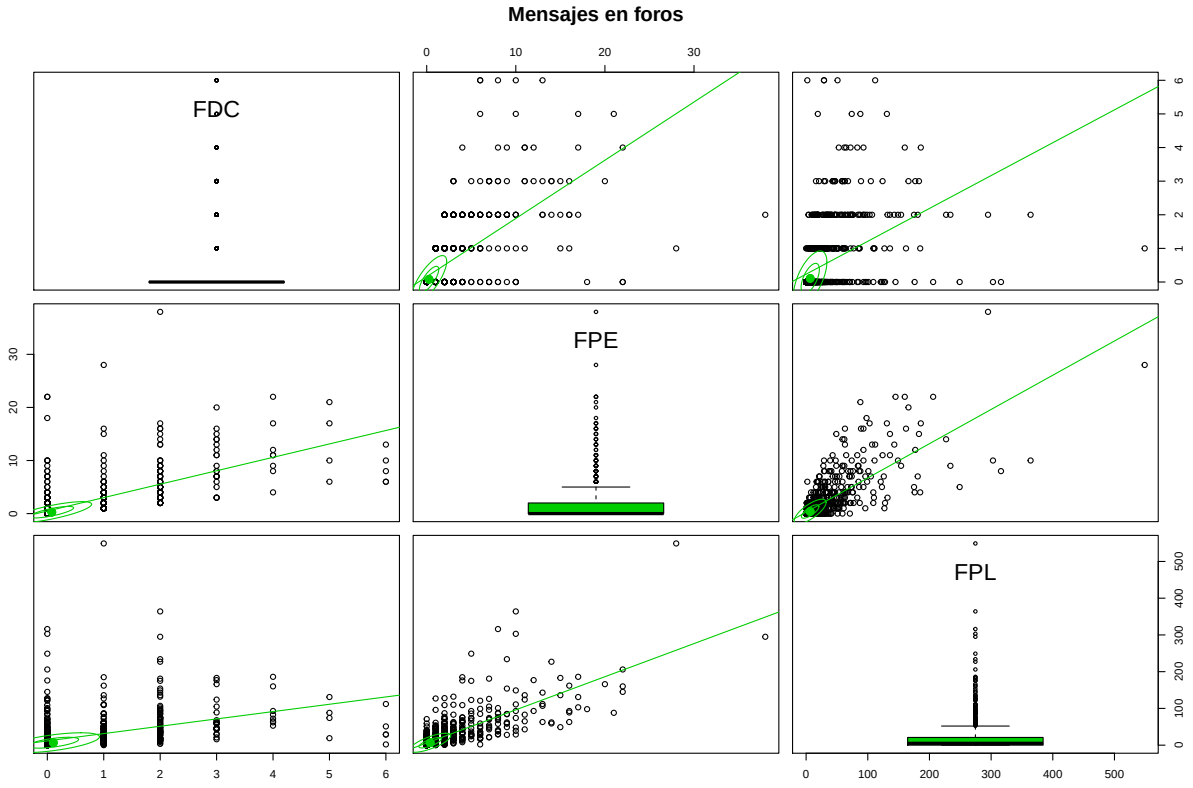


Figura 2.8: Mensajes en foros

Resaltamos aquí la relación entre los mensajes en foros creados (nuevo tema de consulta), enviados (como respuesta a un tema ya propuesto y leídos (consultados). Notar la escala de valores de estos últimos mucho mayor que los dos anteriores.

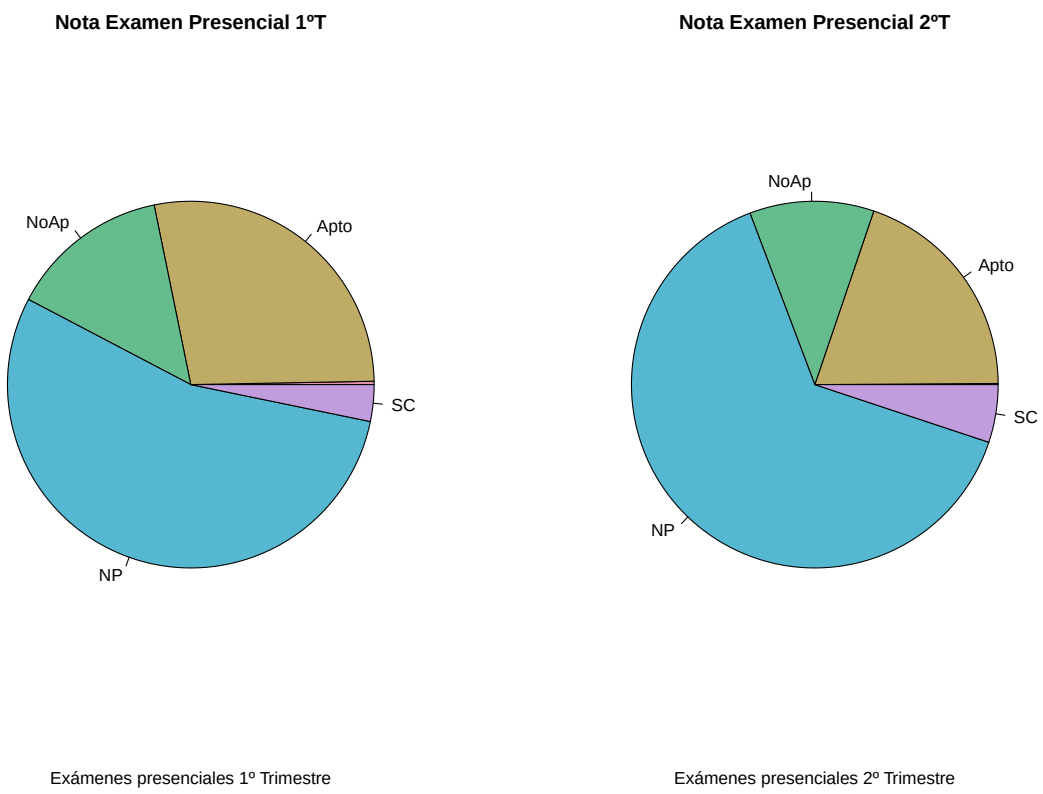


Figura 2.9: Exámenes presenciales 1º, 2º trimestres

En el caso de la realización de exámenes presenciales observamos como en la segunda evaluación el procentaje de presentados es menor que en la evaluación anterior, disminuyendo así el número de “Aptos”.

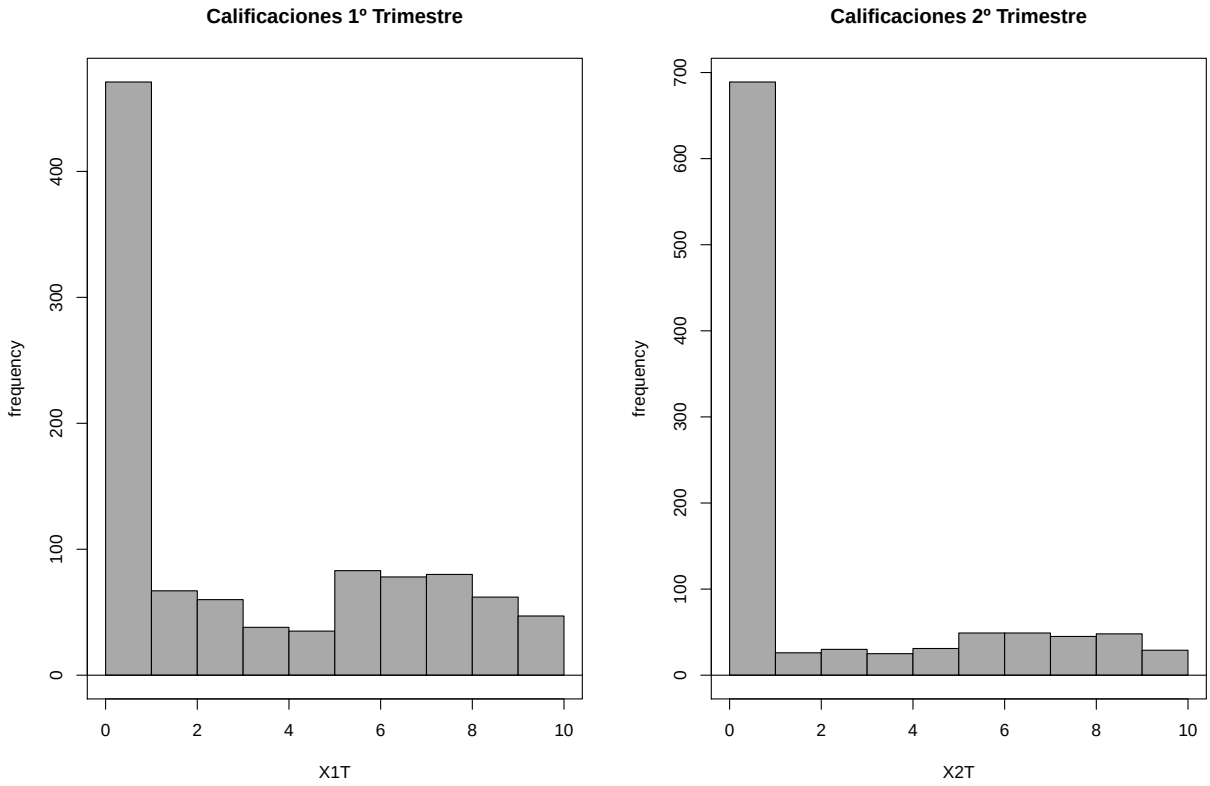


Figura 2.10: Calificaciones 1º, 2º trimestres

En general, las calificaciones obtenidas en el primer trimestre suelen ser mejores que en el segundo, hecho provocado no sólo por la motivación inicial del alumno, si no por la creciente dificultad de las tareas y el necesario aumento de conceptos que el alumno necesita en cada trimestre.

2.5. Informe automático de estadísticas descriptivas

El objetivo de este informe no es otro que el explotar los datos con los que se cuenta según unos requisitos fijados. Aunque es esta una línea de trabajo en la que se pretende en un futuro próximo ir creciendo, en principio se establecen una serie de prioridades a las que intentar dar respuesta.

Por un lado, hay que destacar que la información que se obtiene con este informe es de gran valor para el centro. Sin embargo, debido al **volumen de datos** y al **caracter dinámico** de los mismos, esto es, alumnos que se matriculan en fechas imprevistas, bajas de matrículas, períodos de inactividad del alumnado, picos acusados de trabajo como consecuencia de entregas masivas de tareas, etc, se hace necesario procesar dichos datos en un momento puntual del tiempo de una manera cómoda y precisa(David Masip Rodó, s.f.).

Así se opta por construir en el lenguaje de programación R scripts que nos permitan **automatizar** el proceso(Crawley 2013). Para presentar los resultados de una manera más vistosa y profesional se opta por emplear R Markdown(RStudio, s.f.). Con esto conseguimos generar los informes automáticamente tanto en formato HTML (para consulta vía web) como en formato PDF (para obtener el resultado impreso)(Alexander Borbón A. 2013).

Tras varias sesiones para darle forma a dicho informe se opta por un modelo concreto estándar(Venancio Tomeo Perucha 2009), aunque como se ha dicho, en un futuro no muy lejano se pretende completar y actualizar el modelo inicial.

A continuación se muestra la **salida en formato PDF de un extracto del informe** completo, concretamente se trata de la parte referente a *1º de Bachillerato*. El informe completo incluye cada uno de los tres trimestres más la convocatoria de junio para cada curso (1º, 2º) y a su vez para cada enseñanza (ESA, Bachillerato).

2.5.1. Salida impresa en formato PDF

[Bachillerato - Trimestre 2 - 22/06/2016]

1. MARCO DE DATOS Fuente: Informe de tutores a partir de la plataforma Moodle.

Fecha del fichero: 02/06/2016.

Enseñanza: Bachillerato.

Alumno Activo: Alumno con al menos 1 Tarea Entregada en el trimestre.

Variables: Las variables originales y calculadas, junto con el significado de las mismas, se detallan en la tabla siguiente:

	Variable	Descripción
1	NotaMeSEN	Nota media en Séneca
2	NotaMe	Nota media de las tareas
3	NotaMeAc	Nota media de las tareas del alumnado activo
4	NumTE	Nº tareas entregadas
5	NumTA	Nº tareas aprobadas
6	NumAlu	Nº total de alumnos
7	NumAluAc	Nº alumnos activos
8	PorAluAc	% alumnos activos
9	MeTEAc	Media tareas entregadas por alumno activo
10	PorTEAluAc	% tareas entregadas por alumno activo
11	NumAluPre	Nº alumnos presentados a la prueba presencial
12	NumAluApto	Nº alumnos aptos en la prueba presencial
13	NumAluApro	Nº alumnos aprobados en el trimestre
14	PorAproTot	% aprobados sobre el total de alumnos
15	PorAproAc	% aprobados sobre activos
16	NumTT	Nº total tareas a evaluar
17	NumTTAlu	Nº total tareas a realizar por alumno
18	PorTE	% tareas entregadas respecto al total de tareas a evaluar
19	PorAluAptoPre	% alumnos aptos de los presentados
20	PorAluAptoTot	% alumnos aptos del total de alumnos
21	PorAluPre	% alumnos presentados

2. ENSEÑANZAS En el presente informe las enseñanzas y niveles (cursos) descritos son:

- Bachillerato 1.
- Bachillerato 2.

2.1. BACHILLERATO 1

2.1.1. Aulas

El conjunto total de aulas a considerar en esta enseñanza y nivel está formado por las siguientes:

Tabla 2.10: Aulas Bachillerato 1:

	Aula
1	1º Bach - Biología y Geología [Profesor 1]
2	1º Bach - Cultura Audivisual [Profesor 2]
3	1º Bach - Dibujo Artístico [Profesor 2]
4	1º Bach - Dibujo Técnico [Profesor 3]
5	1º Bach - Dibujo Técnico [Profesor 4]
6	1º Bach - Economía [Profesor 5]
7	1º Bach - Economía [Profesor 6]
8	1º Bach - Filosofía [Profesor 7, Profesor 8]
9	1º Bach - Física y Química [Profesor 9]
10	1º Bach - Francés (Segundo Idioma) [Profesor 10]
11	1º Bach - Francés (Segundo Idioma) [Profesor 11]
12	1º Bach - Francés [Profesor 12]
13	1º Bach - Griego [Profesor 14]
14	1º Bach - Inglés (Segundo Idioma) [Profesor 16]
15	1º Bach - Inglés [Eduardo López]
16	1º Bach - Inglés [Profesor 17]
17	1º Bach - Inglés [Profesor 18]
18	1º Bach - Latín [Profesor 14, Profesor 19]
19	1º Bach - Lengua Castellana y Literatura [Profesor 20]
20	1º Bach - Lengua Castellana y Literatura [Profesor 21]
21	1º Bach - Lengua Castellana y Literatura [Profesor 24]
22	1º Bach - Literatura Universal [Profesor 21]
23	1º Bach - Matemáticas [Profesor 25]
24	1º Bach - Matemáticas Aplicadas a las Ciencias Sociales [Profesor 26]
25	1º Bach - Matemáticas Aplicadas a las Ciencias Sociales [Profesor 27]
26	1º Bach - Matemáticas Aplicadas a las Ciencias Sociales [Profesor 28]
27	1º Bach - Tecnología Industrial [Profesor 29]
28	1º Bach - Volumen [Profesor 2]
29	1º Bach - Ciencias para el Mundo Contemporáneo [Profesor 1]
30	1º Bach - Fundamentos del arte [Profesor 13]

2.1.2. Alumnado

La tabla siguiente muestra por aula un recuento del alumnado, diferenciando entre totales y activos. Por último, se indica el porcentaje de estos últimos respecto al total de alumnos.

Tabla 2.11: Alumnado 2º trimestre:

	Aula	NumAluAc	NumAlu	PorAluAc
1	1º-BG1[P1]	17	44	38.64
2	1º-CA[P2]	3	18	16.67
3	1º-CMC[P1]	2	5	40.00
4	1º-DA1[P2]	6	21	28.57
5	1º-DT1[P3]	8	26	30.77
6	1º-DT1[P4]	2	16	12.50
7	1º-EC1[P5]	23	51	45.10
8	1º-EC1[P6]	14	45	31.11
9	1º-FA[P1]	0	3	0.00
10	1º-FC1[P7,P8]	59	177	33.33
11	1º-FQ[P9]	17	54	31.48
12	1º-FR1P[P1]	4	18	22.22
13	1º-FR1S[P1]	71	148	47.97
14	1º-GR1[P1]	9	37	24.32
15	1º-ING1P[P1]	55	171	32.16
16	1º-ING1S[P1]	1	6	16.67
17	1º-LE1[P2]	37	115	32.17
18	1º-LT1[P1,P1]	15	50	30.00
19	1º-LU1[P2]	31	78	39.74
20	1º-MA1[P2]	43	90	47.78
21	1º-MT1[P2]	23	66	34.85
22	1º-TI1[P2]	15	30	50.00
23	1º-VO[P2]	3	13	23.08

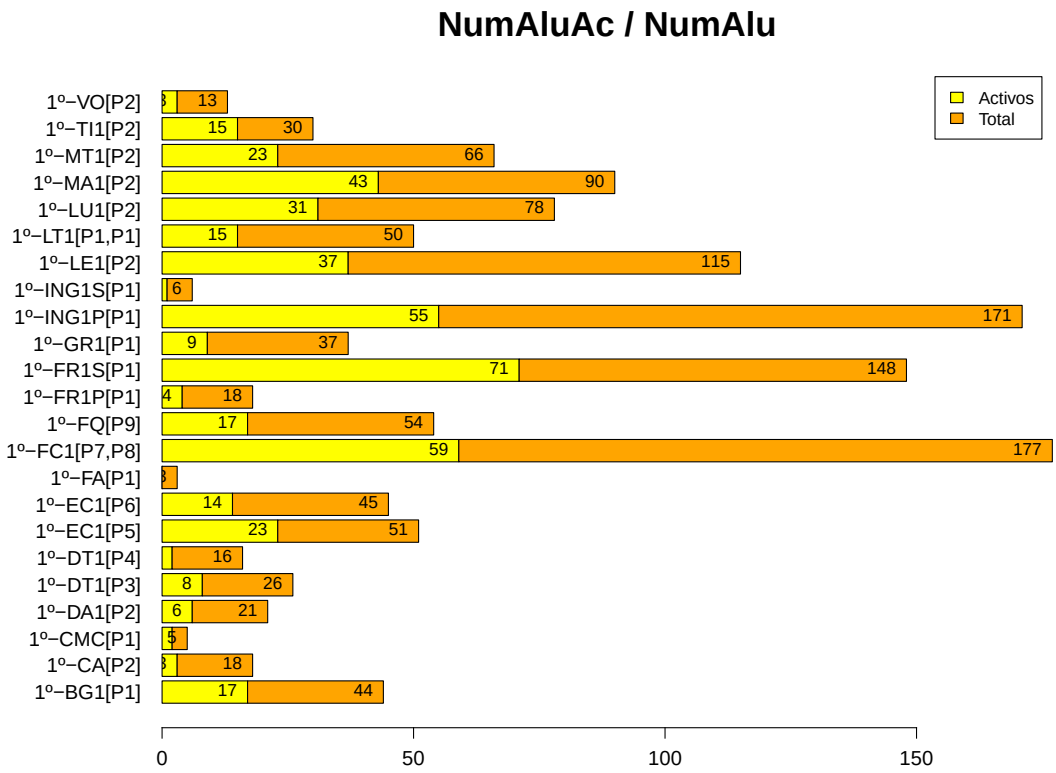


Figura 2.11: Distribución del alumnado

En la gráfica podemos resaltar el hecho de que el aula *FC1[P7,P8]* es compartida por dos profesores y por ello el número de alumnos totales es el máximo de todas, sin embargo, no llega al 50 % de alumnado activo.

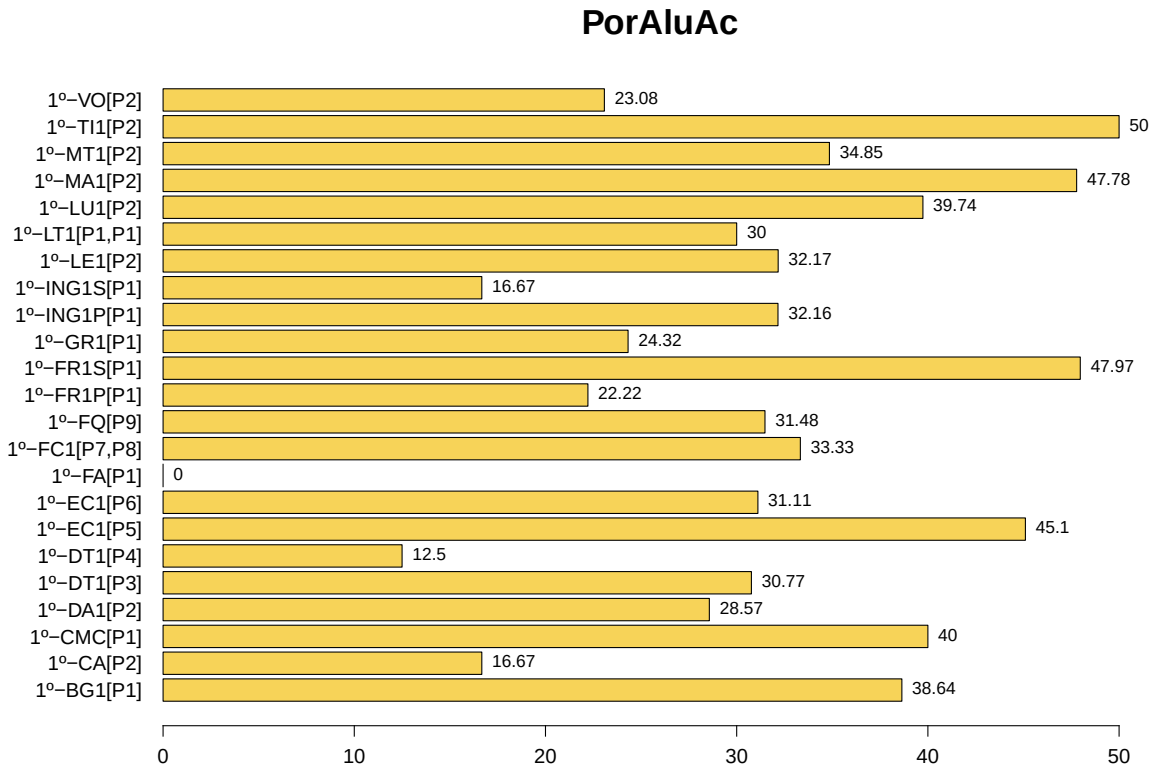


Figura 2.12: Distribución del alumnado en %

La gráfica nos muestra como, en media, el porcentaje de alumnado activo en las aulas es aproximadamente el 30 %, valor considerado por el centro como límite inferior de actividad aceptable. Además ningún valor supera el 50 % de actividad, signo de alerta para intentar al menos no bajar los porcentajes obtenidos en cada aula.

2.1.3. Tareas

Los valores que aparecen en la siguiente tabla hacen referencia al número total de tareas que los alumnos deben realizar, número de tareas entregadas y el porcentaje de tareas que a esta fecha se ha entregado por aula.

Tabla 2.12: Tareas 2º trimestre:

	Aula	NumTTAlu	Nº tareas entregadas	NumTT	PorTE
1	1º-BG1[P1]	6.000000	63	264	23.86
2	1º-CA[P2]	6.000000	15	108	13.89
3	1º-CMC[P1]	5.000000	9	25	36.00
4	1º-DA1[P2]	6.000000	27	126	21.43
5	1º-DT1[P3]	6.000000	41	156	26.28
6	1º-DT1[P4]	6.000000	11	96	11.46
7	1º-EC1[P5]	5.000000	96	255	37.65
8	1º-EC1[P6]	5.000000	60	225	26.67
9	1º-FA[P1]	0.000000	0	0	0.00
10	1º-FC1[P7,P8]	5.000000	238	885	26.89
11	1º-FQ[P9]	6.000000	68	324	20.99
12	1º-FR1P[P1]	5.000000	12	90	13.33
13	1º-FR1S[P1]	3.263514	178	483	36.85
14	1º-GR1[P1]	5.000000	36	185	19.46
15	1º-ING1P[P1]	5.000000	171	855	20.00
16	1º-ING1S[P1]	4.000000	2	24	8.33
17	1º-LE1[P2]	5.000000	124	575	21.57
18	1º-LT1[P1,P1]	5.000000	58	250	23.20
19	1º-LU1[P2]	6.000000	158	468	33.76
20	1º-MA1[P2]	6.000000	172	540	31.85
21	1º-MT1[P2]	6.000000	95	396	23.99
22	1º-TI1[P2]	6.000000	67	180	37.22
23	1º-VO[P2]	6.000000	13	78	16.67

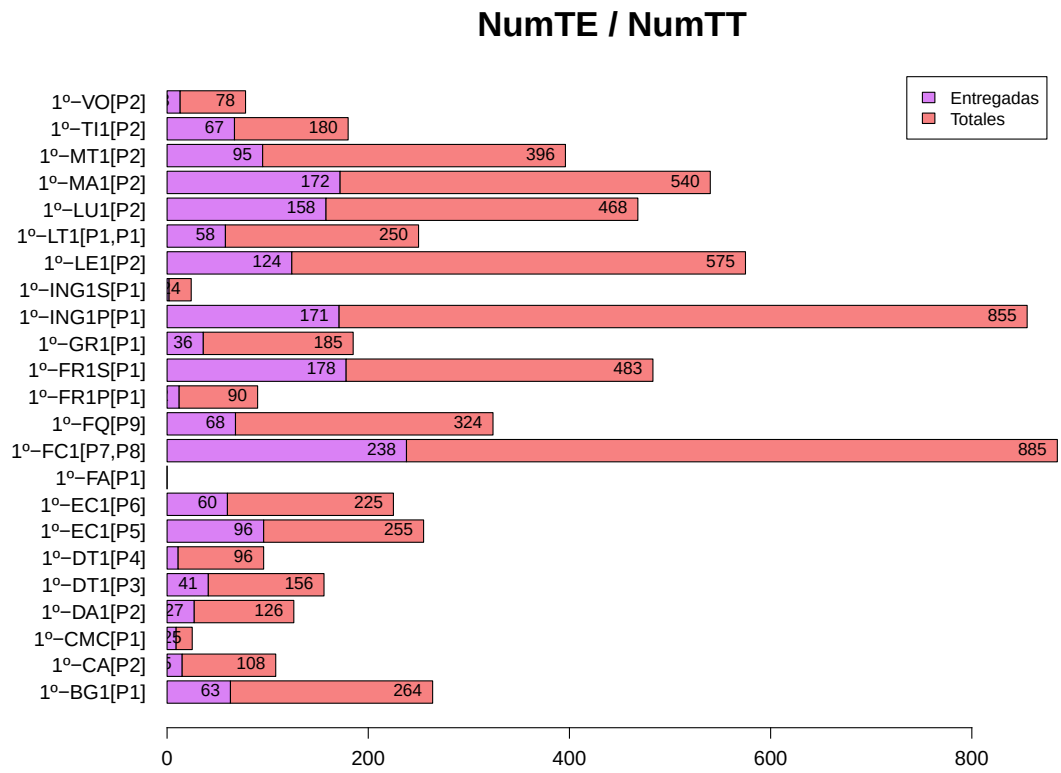


Figura 2.13: Distribución de tareas

Las aulas en las que la diferencia entre las tareas entregadas y tareas totales es mayor, casos *ING1P[P1], FC1[P7,P8]* reflejan tanto la menor actividad del alumnado como la reducción de la carga de trabajo del profesorado asociado según los valores teóricos.

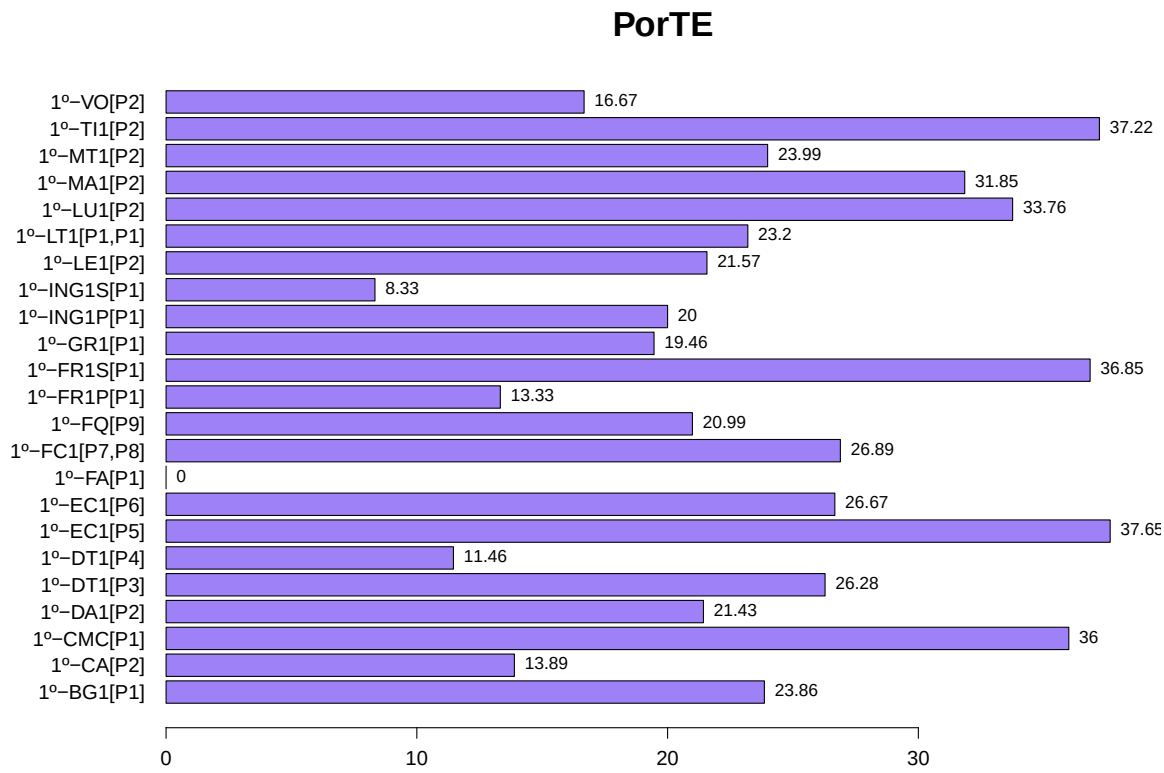


Figura 2.14: Distribución de tareas en %

El porcentaje medio de tareas se sitúa en torno al 20 %, en las aulas donde este valor es inferior, el profesor debe revisar la actividad del alumnado activo para, a ser posible, incrementar dicho porcentaje, puesto que la práctica nos revela que existen alumnos que se conformar con el aprobado y dejan de enviar tareas para subir su calificación, hecho que luego repercute negativamente cuando realizan las pruebas presenciales.

2.1.4. Pruebas Presenciales

Para superar el trimestre, uno de los requisitos necesarios es superar la correspondiente prueba presencial. A continuación, se muestran los resultados referentes a este punto, detallando con respecto al total del alumnado y al número de alumnos aptos.

Tabla 2.13: Pruebas presenciales 2º Trimestre:

	Aula	NumAluApto	NumAluPre	NumAlu	PorAluAptoPre
1	1º-BG1[P1]	8	10	44	80.00
2	1º-CA[P2]	3	5	18	60.00
3	1º-CMC[P1]	2	2	5	100.00
4	1º-DA1[P2]	5	6	21	83.33
5	1º-DT1[P3]	2	9	26	22.22
6	1º-DT1[P4]	2	4	16	50.00
7	1º-EC1[P5]	17	21	51	80.95
8	1º-EC1[P6]	9	11	45	81.82
9	1º-FA[P1]	0	0	3	0.00
10	1º-FC1[P7,P8]	39	58	177	67.24
11	1º-FQ[P9]	7	13	54	53.85
12	1º-FR1P[P1]	1	4	18	25.00
13	1º-FR1S[P1]	35	53	148	66.04
14	1º-GR1[P1]	4	8	37	50.00
15	1º-ING1P[P1]	17	59	171	28.81
16	1º-ING1S[P1]	0	0	6	0.00
17	1º-LE1[P2]	16	30	115	53.33
18	1º-LT1[P1,P1]	8	9	50	88.89
19	1º-LU1[P2]	23	27	78	85.19
20	1º-MA1[P2]	22	34	90	64.71
21	1º-MT1[P2]	15	20	66	75.00
22	1º-TI1[P2]	13	13	30	100.00
23	1º-VO[P2]	3	4	13	75.00

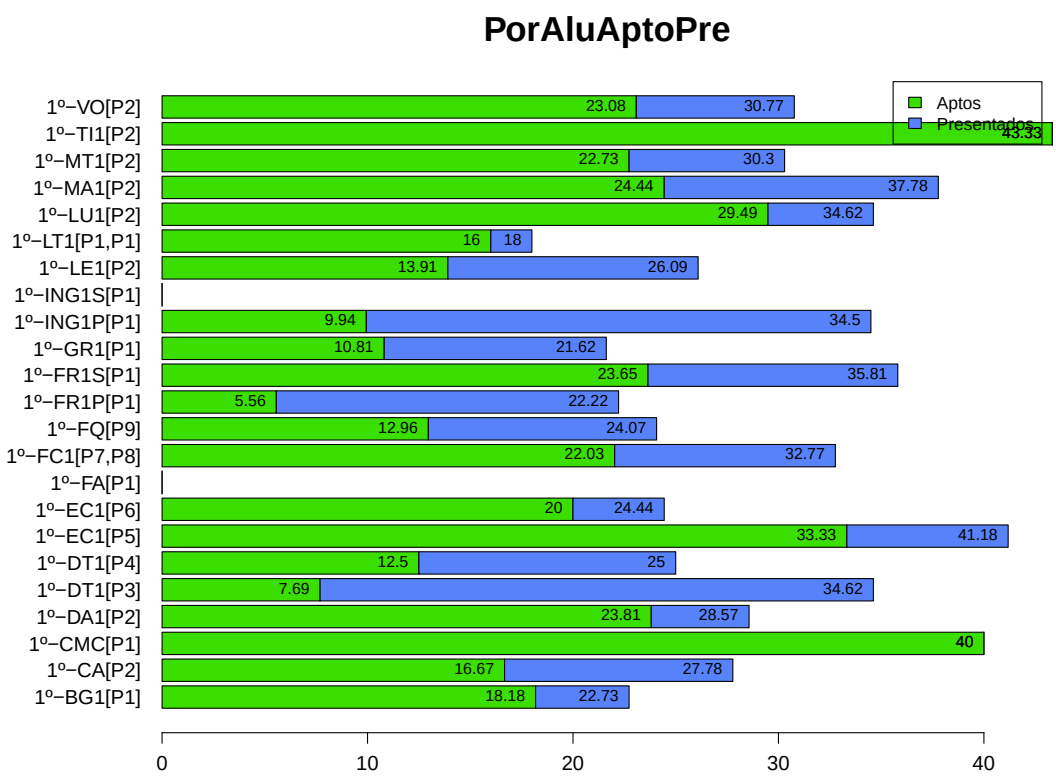


Figura 2.15: Distribución de pruebas presenciales en %

En la mayoría de las aulas se da el resultado esperado por el centro de que más del 50 % de los alumnos presentados a las pruebas presenciales optienen *Apto* en su calificación.

2.1.5. Calificaciones

En relación a las calificaciones (notas) que el alumnado obtiene, por un lado como media ponderada de las Tareas Entregadas y por otro al realizar la prueba presencial, los resultados se muestran en la tabla siguiente:

Tabla 2.14: Calificaciones 2º trimestre:

	Aula	NotaMe	NotaMeAc	NumAluApro	NumAluAc	PorAproTot	PorAproAc
1	1º-BG1[P1]	1.69	4.38	7	17	15.91	41.18
2	1º-CA[P2]	1.18	7.06	3	3	16.67	100.00
3	1º-CMC[P1]	2.99	7.48	2	2	40.00	100.00
4	1º-DA1[P2]	1.57	5.48	5	6	23.81	83.33
5	1º-DT1[P3]	2.00	6.51	2	8	7.69	25.00
6	1º-DT1[P4]	0.97	7.73	2	2	12.50	100.00
7	1º-EC1[P5]	2.48	5.49	15	23	29.41	65.22
8	1º-EC1[P6]	2.01	6.46	7	14	15.56	50.00
9	1º-FA[P1]	0.00	0.00	0	0	0.00	0.00
10	1º-FC1[P7,P8]	1.96	5.89	35	59	19.77	59.32
11	1º-FQ[P9]	1.87	5.95	7	17	12.96	41.18
12	1º-FR1P[P1]	0.88	3.97	1	4	5.56	25.00
13	1º-FR1S[P1]	2.77	5.72	32	71	21.62	45.07
14	1º-GR1[P1]	1.37	5.62	3	9	8.11	33.33
15	1º-ING1P[P1]	1.30	4.03	11	55	6.43	20.00
16	1º-ING1S[P1]	0.50	3.00	0	1	0.00	0.00
17	1º-LE1[P2]	1.43	4.44	14	37	12.17	37.84
18	1º-LT1[P1,P1]	1.61	5.37	9	15	18.00	60.00
19	1º-LU1[P2]	2.56	6.43	24	31	30.77	77.42
20	1º-MA1[P2]	2.41	5.05	14	43	15.56	32.56
21	1º-MT1[P2]	2.07	5.95	13	23	19.70	56.52
22	1º-TI1[P2]	3.48	6.96	12	15	40.00	80.00
23	1º-VO[P2]	1.31	5.69	2	3	15.38	66.67

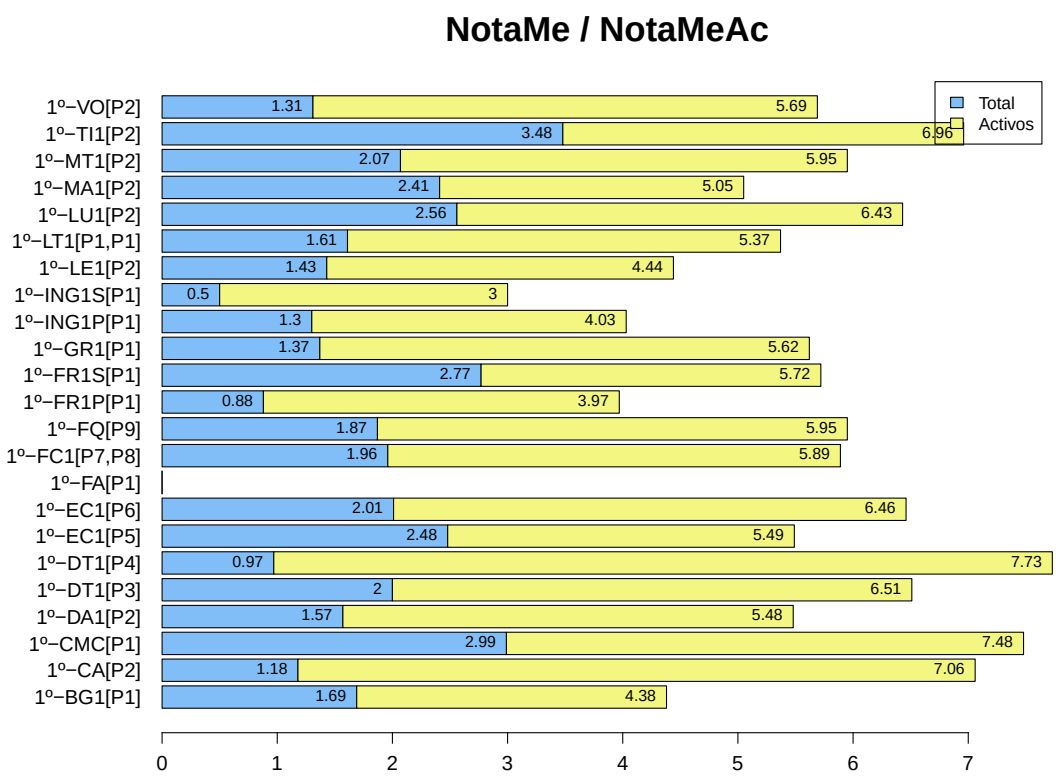


Figura 2.16: Nota media del total de alumnos frente a nota media del alumnado activo

Los resultados obtenidos responden lógicamente a la idea de que la nota media del alumnado activo debe superar en todas las aulas la nota media del total del alumnos. En caso contrario se deben buscar las causas, pues es indicio de error en el calificador de la plataforma.

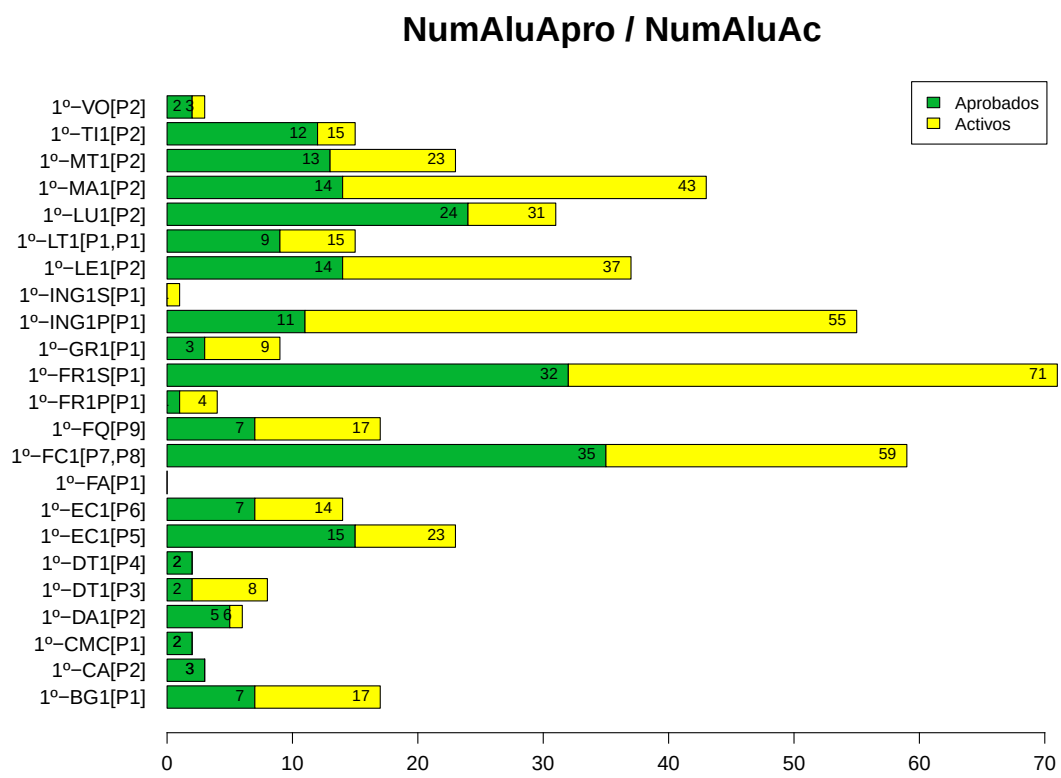


Figura 2.17: Nº de alumnos aprobados frente al total de alumnos activos en %

Tras los resultados expuestos en esta gráfica, el profesor debe localizar qué alumnos activos no han superado la prueba presencial con el fin de orientarles o conocer mejor su situación.

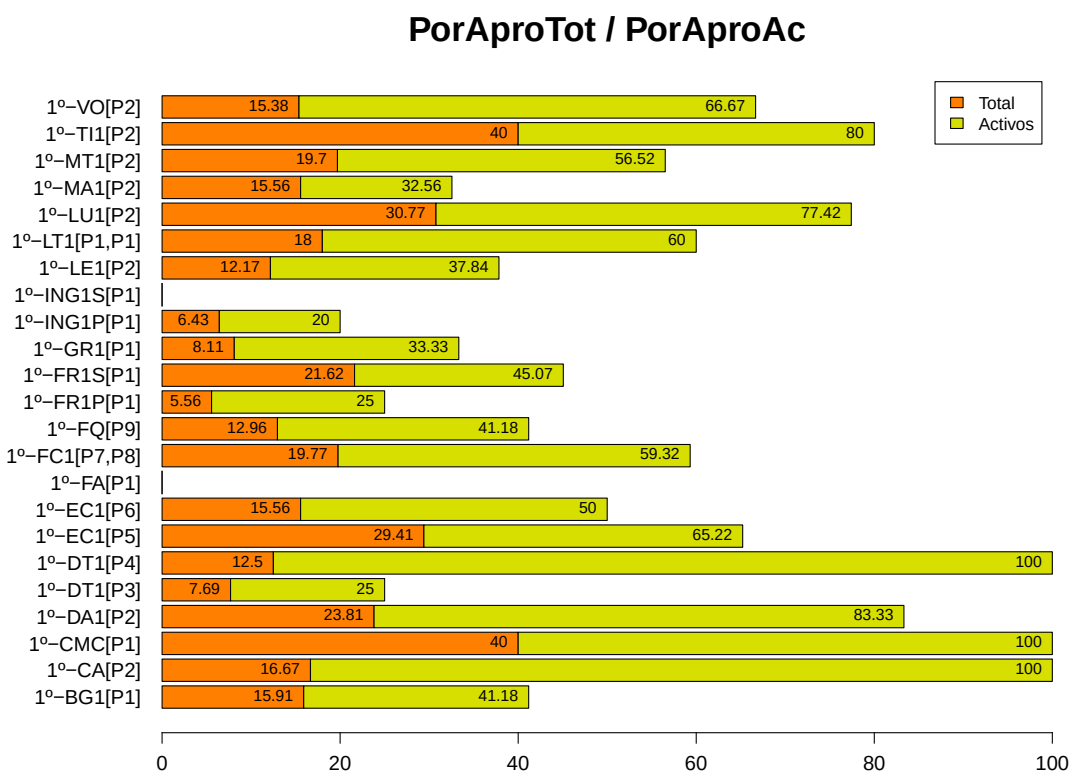


Figura 2.18: Nº de alumnos aprobados frente al total de alumnos activos

El posible abandono en algunas aulas queda evidente en este gráfico, por tanto, es preciso buscar las posibles causas que han dado lugar a ello para mejorar en el futuro.

Capítulo 3

Modelos para predicción

Tras los resultados obtenidos en los informes de estadísticas descriptivas, el centro se plantea un paso más, ahora ya sabemos qué está pasando pero ¿qué pasaría bajo unas condiciones determinadas?

Se pretende por tanto construir varios modelos para la predicción de algunas características (caso cualitativo) y de calificaciones de un trimestre concreto (caso cuantitativo). Hay dos cuestiones en este sentido de las que se quiere obtener respuesta. Por un lado, conocer de antemano qué alumnos causarán baja en su matrícula antes del último plazo de matriculación. También sería interesante obtener una predicción de la calificación del tercer trimestre al comienzo del mismo.

3.1. Modelo de clasificación SVM para predicción de bajas de matrícula

Hasta la fecha, el protocolo de detección, orientación y avisos a posibles alumnos que pudieran no formalizar su matrícula pasada la primera evaluación (mes de febrero) se basaba principalmente en la información particular que cada profesor aportaba de cada uno de sus alumnos. Ello tenía varios invonvenientes, entre otros:

- Coste en el tiempo empleado por cada profesor en gestionar las comunicaciones con su alumnado, vía email o por teléfono, restando así al de otras obligaciones prioritarias, esto es, diseño de actividades online (tareas) y actividades presenciales, atención a consultas de alumnos, procesos de evaluación, etc.
- Coste económico en comunicaciones telefónicas, avisos por correo, personal administrativo, etc.
- Imposibilidad de contactar con determinados alumnos para conocer su situación individual.

Se pretende por tanto construir un modelo de clasificación que permita estimar de forma sistemática qué alumno/s causaría/n baja antes de que termine el fin del proceso de matriculación. De esta forma, se pretende *etiquetar* cada caso y poder así actuar en consecuencia en los que el modelo nos detecte como *posibles bajas*(Mejías 2015).

3.1.1. Breve explicación teórica de la técnica SVM

Dentro de la tarea de clasificación(Suárez 2014), las Máquinas de Vectores Soporte, Máquinas de Soporte Vectorial o en inglés *Support Vector Machine*, en adelante SVM, pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original de los ejemplos de entrada, si éstos son separables o cuasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original. Como se verá más adelante, la búsqueda del hiperplano de separación en estos espacios transformados, normalmente de muy alta dimensión, se hará de forma implícita utilizando las denominadas *funciones kernel*.

Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVMs radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un *margen máximo* a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de *vectores soporte*. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento.

Desde un punto de vista algorítmico, el problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante técnicas estándar de programación cuadrática. La propiedad de convexidad exigida para su resolución garantizan una *solución única*, en contraste con la no unicidad de la solución producida por una red neuronal artificial entrenada con un mismo conjunto de ejemplos.

3.1.2. Metodología propuesta

Aunque podría haberse planteado la opción de regresión logística(Blandón 2012), se ha optado por clasificar empleando la técnica de **Máquinas de Vectores Soporte (SVM)**. La idea es construir un modelo de clasificación con el objetivo de que en cursos posteriores se pueda con antelación conocer qué posibles alumnos pueden causar baja en la matrícula (abandono) y actuar en consecuencia.

A continuación se detalla el proceso seguido hasta obtener el modelo SVM de clasificación:

En primer lugar, partimos de dos ficheros de datos obtenidos en fechas diferentes, esto es, el primero de ellos de principios de febrero y el segundo con fecha de finales de marzo. El fichero con fecha posterior contiene sólo los alumnos que sí son definitivos y por tanto han optado por completar el curso, así los alumnos que aparecen en el fichero más grande, este es el de comienzos de febrero, pueden aún causar baja en su matrícula. Nuestro objetivo se centrará en *etiquetar* cada alumno *caso* como posible baja.

La importancia de obtener este modelo queda justificada por diversas cuestiones:

- De cara al *profesor* permitiría con un mes de antelación planificar más eficientemente el trabajo, estimando con qué porcentaje del alumnado se quedaría una vez pasada la fecha límite de matrícula, gestionando así mejor el tiempo que dedicará a evaluar, responder dudas, etc.
- El tema *administrativo* es también parte implicada aquí. Actualmente durante el mes de febrero se notifica a cada alumno que debe cerrar su matrícula y los plazos para hacerlo. Con un total aproximado de 5000 alumnos el tiempo y dinero para realizar esta labor es bastante considerable.

El primer paso para poder trabajar con los datos consiste en solucionar algunos *problemas de codificación* de ficheros y corregir el nombre de algunas variables que por defecto aparecen con signos de puntuación, caracteres especiales, etc.

Al importar los ficheros, algunas variables no aparecen con el *tipo* correcto, es decir, fechas, factores, etc. Será necesario revisar todas y cada una de ellas convirtiendo al tipo adecuado.

Nos aparecen dos variables de tipo fecha, estas son, *Primer_acceso* y *Ultimo_acceso* que aunque de suma importancia difíciles de tratar puesto que considerarlas por ejemplo de tipo factor implicaría un número demasiado elevado de niveles. Trabajar directamente con su tipo natural, tipo fecha, puede ser también una desventaja pues en ocasiones pueden no ser intuitivas y confundir los resultados que se pueden obtener operando con ellas (sumas y restas). Así, se opta por crear dos nuevas variables etiquetadas como *Días_hasta_primer_acceso* y *Dias_desde_ultimo_acceso* que contienen respectivamente los días que pasan desde el principio de curso hasta que un alumno entra la primera vez en la plataforma virtual de formación y la diferencia de días entre la fecha del fichero de datos y la fecha del último día de acceso a la plataforma.

Observando la primera de ellas *Días_hasta_primer_acceso* vemos que hay algunos casos de alumnos con valor cero en esta variable y sin embargo tienen una calificación de 6 o más en la primera evaluación, se trata de alumnos repetidores que simplemente se presentan a la prueba presencial y por tanto no necesitan acceder a la plataforma puesto que en cursos anteriores realizaron las tareas online con éxito. Es lógica la opción de no considerar estos casos pues podrían distorsionar el modelo de predicción.

Cruzando los datos de ambos ficheros obtenemos la relación real de alumnos que se han dado de *baja* y cuáles siguen en el curso. De esta forma tenemos etiquetados todos los casos del fichero inicial (antes de la posibilidad de causar baja), para ello creamos una nueva variable *Baja* de tipo factor con los valores “SI”, “NO” según el caso.

Para elegir qué variables usar analizamos la correlación entre las variables candidatas, en este caso consideramos importantes *Dias_desde_ultimo_acceso* y *NOTA_1T*:

Analizamos la correlación entre las variables que consideramos aportan información a la actividad del primer trimestre para construir el modelo SVM:

```
##                          Dias.hasta.Primer.acceso Dias.desde.Ultimo.acceso
## Dias.hasta.Primer.acceso               1.00000000              -0.12142197
## Dias.desde.Ultimo.acceso              -0.12142197               1.00000000
## TE.1T                                 -0.15168196              -0.44146460
## TA.1T                                 -0.15427560              -0.42913367
## X1T.SENECA                            -0.20273320              -0.42602816
## CE                                    -0.04039162              -0.10188701
## CR                                    -0.11893644              -0.05335165
## FPE                                   -0.13966105              -0.27820020
## FDC                                   -0.12293633              -0.25725617
## FPL                                   -0.14468306              -0.29943885
##                                TE.1T      TA.1T X1T.SENECA          CE
## Dias.hasta.Primer.acceso -0.15168196 -0.1542756 -0.2027332 -0.04039162
## Dias.desde.Ultimo.acceso -0.44146460 -0.4291337 -0.4260282 -0.10188701
## TE.1T                     1.00000000  0.9666295  0.8623560  0.17674818
## TA.1T                     0.96662947  1.0000000  0.8710067  0.17744384
## X1T.SENECA                0.86235600  0.8710067  1.0000000  0.20341471
## CE                        0.17674818  0.1774438  0.2034147  1.00000000
## CR                        0.09830275  0.1049390  0.1717204  0.65580535
## FPE                       0.41098330  0.4209819  0.4205456  0.10420426
## FDC                       0.31951617  0.3297304  0.3077780  0.10833489
## FPL                       0.43712198  0.4323798  0.4408825  0.17777641
##                                  CR         FPE         FDC         FPL
## Dias.hasta.Primer.acceso -0.11893644 -0.13966105 -0.12293633 -0.14468306
## Dias.desde.Ultimo.acceso -0.05335165 -0.27820020 -0.25725617 -0.29943885
## TE.1T                     0.09830275  0.41098330  0.31951617  0.43712198
## TA.1T                     0.10493901  0.42098190  0.32973043  0.43237978
## X1T.SENECA                0.17172037  0.42054557  0.30777802  0.44088253
## CE                        0.65580535  0.10420426  0.10833489  0.17777641
## CR                        1.00000000  0.03428628  0.07176839  0.06653918
## FPE                       0.03428628  1.00000000  0.73471099  0.72952620
## FDC                       0.07176839  0.73471099  1.00000000  0.43970045
## FPL                       0.06653918  0.72952620  0.43970045  1.00000000
```

Una vez elegidas las variables x = *Dias_desde_ultimo_acceso*, y = *NOTA_1T*, *grupo* = *Baja* construimos un nuevo data.frame de la forma *datos<- data.frame(x,y,grupo)*.

Representamos gráficamente y observamos que excepto en la zona central, podemos ver dos grupos diferenciados, a medida que crece la variable *Dias_desde_ultimo_acceso*.

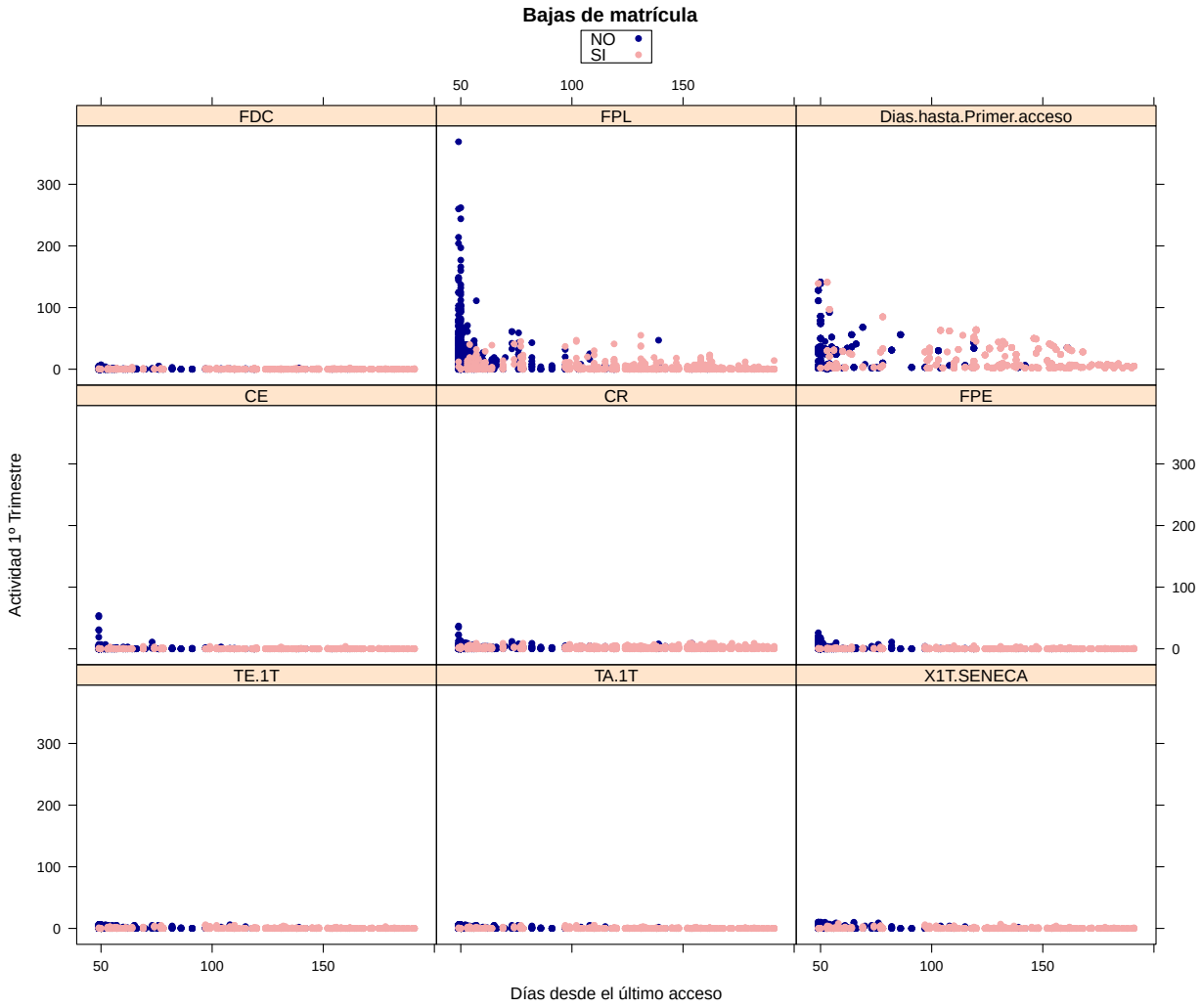


Figura 3.1: Bajas de matrícula

Necesitamos una muestra de entrenamiento, para ello consideramos el 70 % de los datos y a través de la función *svm* de la librería *e1071* construimos el modelo. Elegimos el *kernel* de tipo radial debido al comportamiento de los valores centrales de los datos. El modelo que obtenemos es el siguiente:

```
##
## Call:
## svm(formula = Baja ~ ., data = m.entrena, method = "C-classification",
##     kernel = "radial", cost = 10, gamma = 0.1, na.action = na.omit)
##
##
## Parameters:
```

```
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  10
##       gamma:  0.1
##
## Number of Support Vectors:  338
##
##  ( 174 164 )
##
##
## Number of Classes:  2
##
## Levels:
##  NO SI
```

3.1.3. Evaluación del modelo

Para evaluar el modelo consideramos el 30 % de casos restantes calculando la **Tabla de confusión** y la representación gráfica correspondiente:

```
## Muestra: (20 primeros casos)
```

```
##    Baja Prediccion
## 9    NO         NO
## 11   NO         NO
## 12   NO         NO
## 13   NO         NO
## 15   NO         NO
## 16   NO         NO
## 17   NO         NO
## 18   NO         NO
## 20   NO         NO
## 22   NO         NO
## 28   NO         NO
## 32   NO         NO
## 35   NO         NO
## 38   NO         NO
## 39   NO         NO
## 42   NO         NO
## 44   SI         SI
## 46   SI         SI
## 47   SI         SI
## 48   SI         SI
```

```
## Tabla de confusión:

##
##       NO  SI
##   NO 387  54
##   SI  76 363

## Tasa de acierto =85.22727%

## Error total de clasificación = 14.77273%

## Error No detectados como baja = 17.31207%

## Error dados como baja = 12.30068%
```

3.1.4. Resultados y conclusiones

La siguiente figura explica gráficamente los resultados obtenidos a partir del modelo SVM que se ha considerado.

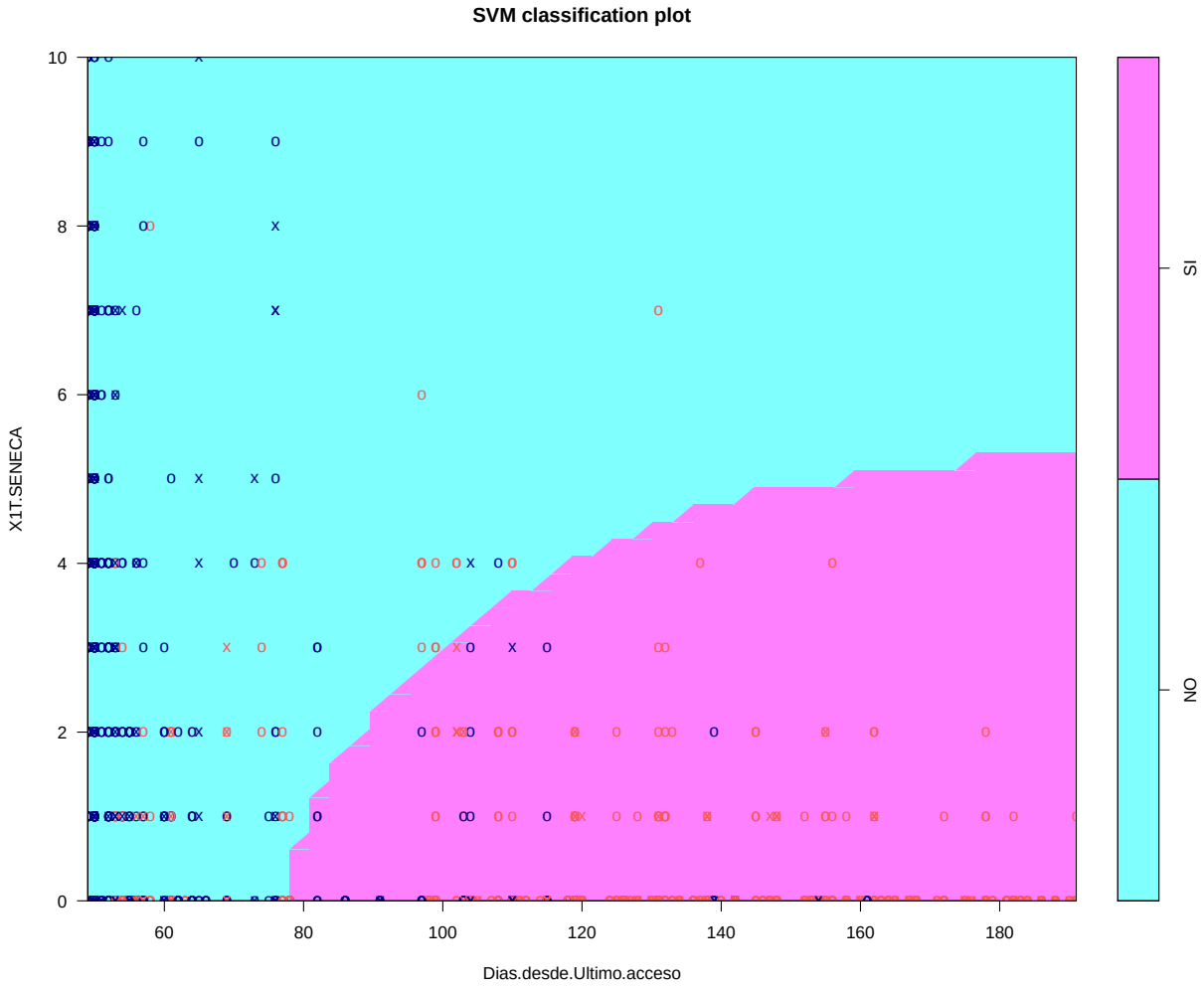


Figura 3.2: Clasificación por SVM

El significado de la misma es el siguiente:

Aparecen dos zonas claramente diferenciadas como consecuencia de la predicción de los dos posibles valores (SI, NO) de la variable *BAJA* de manera que:

- La zona celeste (color claro) es la considerada bajo el modelo como NO baja, es decir, región de predicción de alumnos que no causarían baja.
- La zona rosa (color oscuro) contrariamente indica la región de predicción bajo el modelo del alumnado que SÍ causaría baja.

Los círculos y cruces representan los casos bien clasificados y mal clasificados respectivamente, así:

- Los círculos y cruces azules (oscuros) representan los casos de alumnos que realmente no han causado baja.
- Los círculos y cruces rojos (claros) indican los casos de aquellos alumnos que efectivamente sí se dieron de baja.
- Las cruces (de ambos colores) representan los vectores soporte.
- Los círculos (de ambos colores) representan los elementos que no son vectores soporte.

Aunque la tabla de confusión (evaluación) del modelo nos indicará los casos bien y mal clasificados, así como cuántos de cada subtipo (4 posibles casos), el modelo no debe cometer mucho error en clasificar como no bajas aquellos casos que realmente sí causarán baja.

Fijándonos en las dos regiones y más concretamente en la frontera de separación de las mismas, la **regla de decisión** sería:

```
## Días a partir de los que avisar al alumno = 82
```

```
## Nota mínima para no tener que avisar al alumno = 4
```

De esta forma las conclusiones serían que la calificación no es tan relevante como sí parece serlo el número de días desde el último acceso.

3.2. Modelo de regresión lineal para la estimación de calificaciones

No cabe duda que la valoración final del éxito alcanzado al final de curso queda reflejado en las calificaciones del alumnado. En este sentido, el centro se plantea obtener una predicción de la calificación final de junio con los datos disponibles al final del segundo trimestre y poder así no sólo estimar dichas calificaciones, sino de cara al profesorado hacer balance y planificar la carga de trabajo que tendría una vez pasada la fecha durante el mes de julio.

El *segundo curso de Bachillerato* es el que se ha tomado como referencia y sobre el que se aplicará la técnica MRLM, por la importancia académica que supone para el alumno aprobar en la convocatoria ordinaria de mayo/junio o por el contrario realizar la recuperación en la convocatoria extraordinaria de Septiembre.

Por el hecho de considerar como variable a predecir la calificación final de junio y ser esta de tipo numérico y continua, se propone usar el Modelo de Regresión Lineal Múltiple.

3.2.1. Breve explicación teórica de la técnica MRLM

Mediante un modelo de regresión lineal múltiple (MRLM) tratamos de explicar el comportamiento de una determinada variable que denominaremos variable a explicar, variable endógena o variable dependiente, (y representaremos con la letra Y) en función de un conjunto de k variables explicativas X_0, X_1, ..., X_k mediante una relación de dependencia lineal (suponiendo $X_0 = 1$):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon \; ;$$

siendo ϵ el término de perturbación o error.

Para determinar el modelo anterior, es necesario hallar (estimar) el valor de los coeficientes β_0, β_1, ..., β_k. La linealidad en parámetros posibilita la interpretación correcta de los parámetros del modelo. Los parámetros miden la intensidad media de los efectos de las variables explicativas sobre la variable a explicar y se obtienen al tomar las derivadas parciales de la variable a explicar respecto a cada una de las variables explicativas.

Nuestro objetivo es asignar valores numéricos a los parámetros β_0, β_1, ..., β_k. Es decir, trataremos de estimar el modelo de manera que, los valores ajustados de la variable endógena resulten tan próximos a los valores realmente observados como sea posible.

A fin de poder determinar las propiedades de los estimadores obtenidos al aplicar distintos métodos de estimación y realizar diferentes contrastes, hemos de especificar un conjunto de hipótesis sobre el MRLM que hemos formulado. Existen tres grupos de hipótesis siguientes: *las hipótesis sobre el término de perturbación, las hipótesis sobre las variables explicativas, y las hipótesis sobre los parámetros del modelo.*

3.2.2. Metodología propuesta

Nuestro objetivo es hacer una estimación de qué alumnos pueden aprobar en la convocatoria ordinaria de mayo/junio y hacer un pronóstico de su calificación en dicha convocatoria. Para ello, debemos previamente **filtrar** los datos fuente atendiendo a los requisitos necesarios para este curso y etapa educativa, puesto que un alumno que no los cumpliese no podría aprobar en la convocatoria ordinaria y pasaría directamente a la extraordinaria de Septiembre:

- Eliminar casos de alumnos que en alguna evaluación tienen el valor cero en su calificación.
- Eliminar casos de alumnos que tienen tanto la primera como la segunda evaluación suspensas.

Para una primera perspectiva de qué reflejan los datos disponibles, se representarán gráficamente. Aunque con ello se pueden observar comportamientos o tendencias, será necesario estudiar las correlaciones entre las variables con el objetivo de construir un modelo apropiado.

En nuestro caso, las variables de tipo cuantitativo con las que contamos para construir un MRLM son:

- CE : Número de correos enviados por el alumno a otros miembros del centro (alumnos y/o profesores).
- CR : Número de correos recibidos por el alumno de otros miembros del centro (alumnos y/o profesores).
- FDC : Número de discusiones o hilos creados en foros.
- FPE : Número de mensajes (posts) enviados en foros.
- FPL : Número de mensajes (posts) leídos en foros.
- X1T : Calificación del primer trimestre.
- X2T : Calificación del segundo trimestre.
- X3T : Calificación del tercer trimestre.
- MEDIA.JUNIO : Calificación media final en junio.

Así nuestra variable dependiente Y será MEDIA.JUNIO y el resto serán tomadas como variables independientes o explicativas.

Puesto que la estimación de la variable Y (MEDIA.JUNIO) la queremos realizar con los datos disponibles una vez finalizado el segundo trimestre, debemos descartar la variable X3T puesto que en ese momento del tiempo no la tendremos.

Concluyendo, nuestro **modelo inicial** (el más completo) será de la forma:

$$\text{MEDIA.JUNIO} = \beta_0 + \beta_1 \text{CE} + \beta_2 \text{CR} + \beta_3 \text{FDC} + \beta_4 \text{FPE} + \beta_5 \text{FPL} + \beta_6 \text{X1T} + \beta_7 \text{X2T}$$

Para explorar las relaciones entre todas las parejas de variables, en particular la relación de Y (*MEDIA.JUNIO*) con cada una de las variables independientes visualizamos los diagramas de dispersión de cada par de variables y estudiamos las correlaciones entre todas las variables.

```
##                      CE          CR           FDC         FPE       FPL
## CE           1.000000000  0.75324319  0.001175049 0.01555996 0.1127993
## CR           0.753243186  1.00000000 -0.040166624 0.02830673 0.1166061
## FDC          0.001175049 -0.04016662  1.000000000 0.65892029 0.4126855
## FPE          0.015559955  0.02830673  0.658920288 1.00000000 0.7206165
## FPL          0.112799305  0.11660605  0.412685476 0.72061655 1.0000000
## X1T          0.121999353  0.13249444  0.085894249 0.17125915 0.1673100
## X2T          0.138625952  0.10747263  0.092325987 0.19296505 0.1863293
## MEDIA.JUNIO 0.160664584  0.14364045  0.090812308 0.17827158 0.1835787
##                     X1T        X2T MEDIA.JUNIO
## CE           0.12199935 0.13862595  0.16066458
## CR           0.13249444 0.10747263  0.14364045
## FDC          0.08589425 0.09232599  0.09081231
## FPE          0.17125915 0.19296505  0.17827158
## FPL          0.16730998 0.18632927  0.18357870
## X1T          1.00000000 0.59114647  0.84094705
## X2T          0.59114647 1.00000000  0.87459190
## MEDIA.JUNIO 0.84094705 0.87459190  1.00000000
```

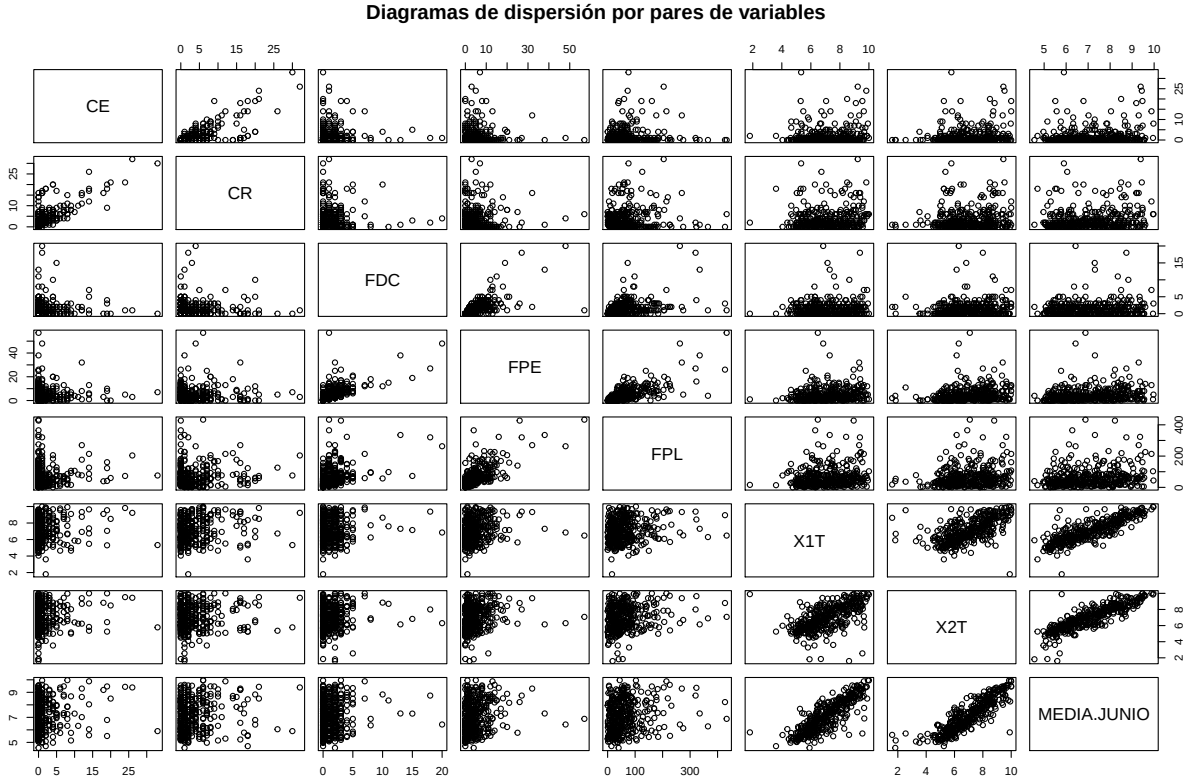


Figura 3.3: Diagramas de dispersión por pares de variables

Vemos que las variables que están fuertemente correlacionadas (como era de esperar) con la

variable MEDIA.JUNIO son X1T, X2T. El resto se correlacionan en menor grado con esta y de forma más o menos similar. Veamos gráficamente de una forma más amplia la relación entre las tres variables que presentan mayor grado de correlación:

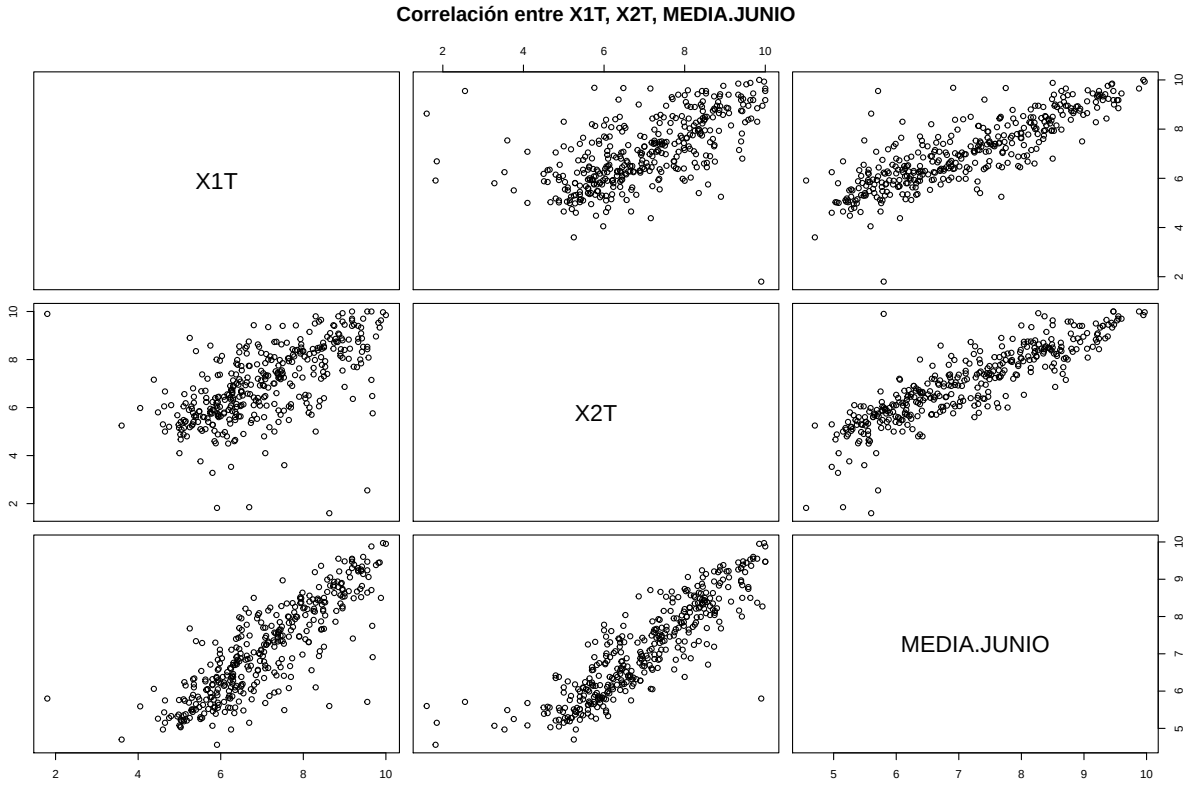


Figura 3.4: Correlación entre X1T, X2T, MEDIA.JUNIO

Llegados a este punto construimos el **modelo completo** con la función *lm* de R y lo resumimos con el objetivo de extraer conclusiones:

```
mod.completo <- lm(MEDIA.JUNIO~CE+CR+FDC+FPE+FPL+X1T+X2T, data=datos)
summary(mod.completo)
```

```
##
## Call:
## lm(formula = MEDIA.JUNIO ~ CE + CR + FDC + FPE + FPL + X1T +
##     X2T, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12580 -0.20860  0.00283  0.22397  0.96916
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5842369  0.0996723   5.862 1.03e-08 ***
## CE           0.0050119  0.0067538   0.742    0.459
## CR           0.0008674  0.0055020   0.158    0.875
## FDC          0.0078171  0.0106122   0.737    0.462
## FPE         -0.0071703  0.0050462  -1.421    0.156
## FPL          0.0001881  0.0004046   0.465    0.642
## X1T          0.4505901  0.0160190  28.128  < 2e-16 ***
## X2T          0.4772167  0.0146199  32.642  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3429 on 362 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9257
## F-statistic: 657.7 on 7 and 362 DF,  p-value: < 2.2e-16
```

La salida nos informa de que tanto la ordenada en el origen β_0 como β_6 y β_7 son significativos, estos últimos fruto de la fuerte correlación con MEDIA.JUNIO. Vemos además que R^2=0.9271057, lo que nos indica el buen ajuste de este modelo.

El hecho de que en el modelo anterior nos aparezcan sólo tres coeficientes de regresión marcados significativamente de los siete posibles nos lleva a plantear la cuestión de qué variables podríamos suprimir de dicho modelo para simplificarlo sin sacrificar mucho el grado de ajuste. Para esto realizamos un estudio de **selección de variables** por ejemplo hacia atrás/adelante (*backward/forward*) atendiendo al criterio AIC. El resultado es el siguiente:

```
##
## Direction:  backward/forward
## Criterion:  AIC
##
## Start:  AIC=-784.07
## MEDIA.JUNIO ~ CE + CR + FDC + FPE + FPL + X1T + X2T
##
##        Df Sum of Sq     RSS     AIC
## - CR    1     0.003  42.574 -786.04
## - FPL   1     0.025  42.596 -785.84
## - FDC   1     0.064  42.634 -785.51
## - CE    1     0.065  42.635 -785.50
## <none>               42.571 -784.07
## - FPE   1     0.237  42.808 -784.01
## - X1T   1    93.045 135.616 -357.36
## - X2T   1   125.298 167.869 -278.42
##
## Step:  AIC=-786.04
## MEDIA.JUNIO ~ CE + FDC + FPE + FPL + X1T + X2T
```

```
##
##        Df Sum of Sq     RSS     AIC
## - FPL   1     0.026  42.599 -787.81
## - FDC   1     0.062  42.635 -787.50
## - CE    1     0.197  42.770 -786.34
## <none>               42.574 -786.04
## - FPE   1     0.236  42.809 -786.00
## + CR    1     0.003  42.571 -784.07
## - X1T   1    93.580 136.154 -357.90
## - X2T   1   125.502 168.075 -279.96
##
## Step:  AIC=-787.81
## MEDIA.JUNIO ~ CE + FDC + FPE + X1T + X2T
##
##        Df Sum of Sq     RSS     AIC
## - FDC   1     0.053  42.653 -789.35
## - CE    1     0.221  42.820 -787.90
## <none>               42.599 -787.81
## - FPE   1     0.250  42.849 -787.65
## + FPL   1     0.026  42.574 -786.04
## + CR    1     0.003  42.596 -785.84
## - X1T   1    93.692 136.292 -359.52
## - X2T   1   125.669 168.268 -281.54
##
## Step:  AIC=-789.35
## MEDIA.JUNIO ~ CE + FPE + X1T + X2T
##
##        Df Sum of Sq     RSS     AIC
## - FPE   1     0.213  42.866 -789.51
## - CE    1     0.220  42.873 -789.45
## <none>               42.653 -789.35
## + FDC   1     0.053  42.599 -787.81
## + FPL   1     0.018  42.635 -787.50
## + CR    1     0.001  42.652 -787.36
## - X1T   1    93.654 136.307 -361.48
## - X2T   1   125.630 168.283 -283.51
##
## Step:  AIC=-789.51
## MEDIA.JUNIO ~ CE + X1T + X2T
##
##        Df Sum of Sq     RSS     AIC
## - CE    1     0.227  43.092 -789.56
## <none>               42.866 -789.51
## + FPE   1     0.213  42.653 -789.35
## + FPL   1     0.055  42.810 -787.99
```

```
## + FDC   1     0.016  42.849 -787.65
## + CR    1     0.001  42.865 -787.51
## - X1T   1    93.502 136.368 -363.32
## - X2T   1   126.136 169.002 -283.93
##
## Step:  AIC=-789.56
## MEDIA.JUNIO ~ X1T + X2T
##
##        Df Sum of Sq     RSS     AIC
## <none>               43.092 -789.56
## + CE    1     0.227  42.866 -789.51
## + FPE   1     0.219  42.873 -789.45
## + CR    1     0.139  42.953 -788.76
## + FPL   1     0.037  43.055 -787.88
## + FDC   1     0.018  43.074 -787.71
## - X1T   1    94.201 137.293 -362.81
## - X2T   1   127.909 171.001 -281.58


##
## Call:
## lm(formula = MEDIA.JUNIO ~ X1T + X2T, data = datos)
##
## Coefficients:
## (Intercept)          X1T          X2T
##      0.5923       0.4504       0.4764
```

Comprobamos que el **modelo simplificado** obtenido ofrece un grado de ajuste muy similar al modelo completo, el cual se resume en:

```
##
## Call:
## lm(formula = MEDIA.JUNIO ~ X2T + X1T, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12917 -0.20659  0.00343  0.22557  1.01540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.59232    0.09864   6.005  4.6e-09 ***
## X2T          0.47636    0.01443  33.005  < 2e-16 ***
## X1T          0.45035    0.01590  28.324  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3427 on 367 degrees of freedom
## Multiple R-squared:  0.9262, Adjusted R-squared:  0.9258
## F-statistic:  2303 on 2 and 367 DF,  p-value: < 2.2e-16
```

3.2.3. Evaluación del modelo

Si deseamos comparar los dos modelos propuestos (completo y simplificado) la información necesaria se encuentra en las tablas ANOVA de cada modelo. Para hacer la comparación entre los modelos se utiliza la instrucción de R *anova(modelo más sencillo, modelo más complejo)* (Zoritza 2008),(M. 2008), la cual permite comparar dos modelos anidados a través de una prueba F.

```
## Analysis of Variance Table
##
## Model 1: MEDIA.JUNIO ~ X2T + X1T
## Model 2: MEDIA.JUNIO ~ CE + CR + FDC + FPE + FPL + X1T + X2T
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    367 43.092
## 2    362 42.571  5   0.52168 0.8872 0.4897
```

En este caso el contraste de hipótesis es:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: Algún $\beta_i <> 0$, i = 1, . . . , 5.

El valor de Pr(>F)=0.4896634 nos indica que no tenemos evidencias para rechazar H_0 y nos confirma la bondad del ajuste del modelo simplificado.

Si representamos gráficamente un diagrama de dispersión para este modelo, podemos ver tanto la tendencia lineal como la estimada bondad de ajuste que, en principio, parece que podremos obtener.

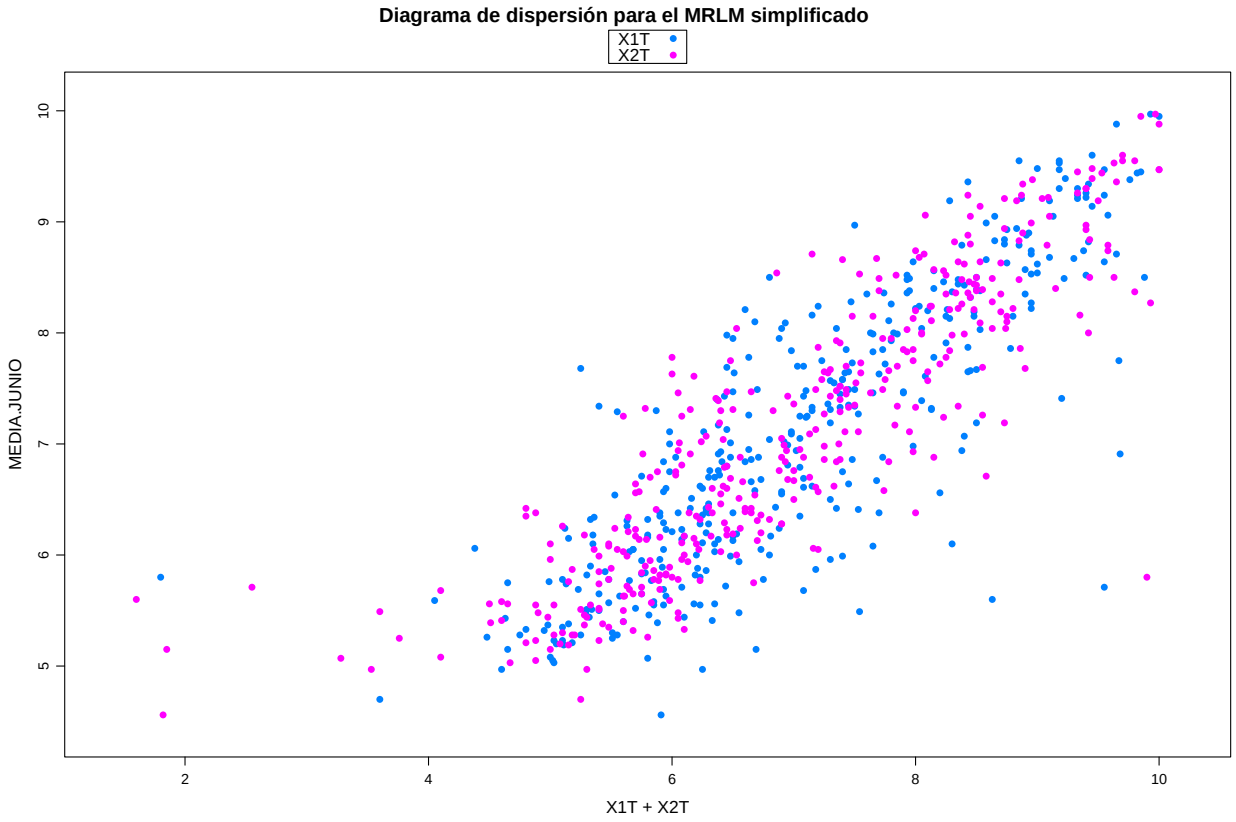


Figura 3.5: Diagrama de dispersión para el MRLM simplificado

El ajuste y el análisis de un modelo lineal se sustentan en cuatro suposiciones básicas:

- La relación entre las variables es lineal, lo cual puede ser chequeado con el diagrama de dispersión de los datos.
- Los errores (residuos) siguen una distribución normal.
- Las varianzas de los errores (residuos) son iguales (es decir los errores son HOMOCEDÁSTICOS).
- Los errores (residuos) son independientes.

Es necesario entonces preguntarse si estas suposiciones se cumplen. R dispone de cuatro representaciones gráficas que nos ayudan a comprobar visualmente estos supuestos. Veámoslas:

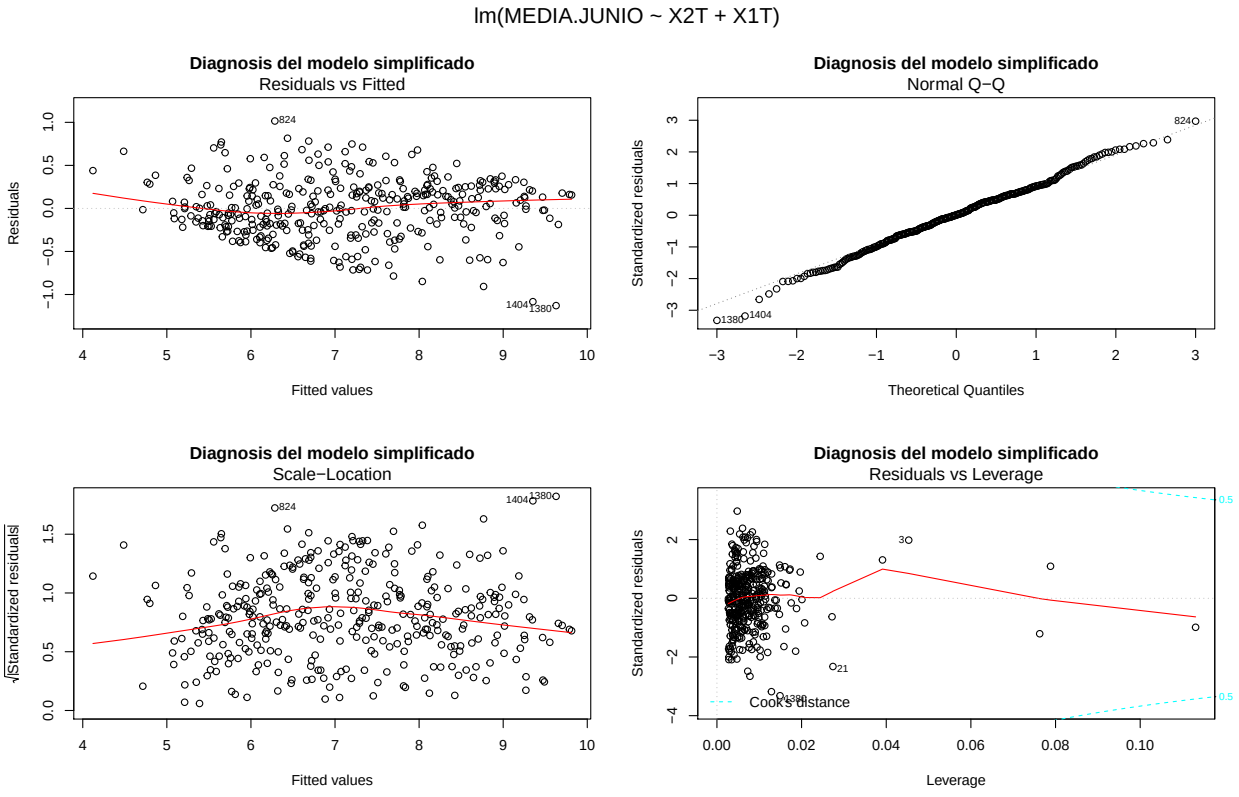


Figura 3.6: Diagnosis del modelo simplificado

Las líneas más o menos rectas representadas son indicativas del cumplimiento de cada hipótesis.

Aunque estos análisis gráficos son muy útiles, a veces es conveniente complementarlos con un contraste de normalidad más formal(Segundo 2013). Exixte una librería en R, llamada **gvlma** (de *Global Validation of Linear Models Assumptions*) que, actuando sobre un modelo producido con *lm* realiza una serie de contrastes sobre las hipótesis del modelo, y nos resume en la respuesta si esas condiciones se verifican. En nuestro caso:

```
##
## Call:
## lm(formula = MEDIA.JUNIO ~ X2T + X1T, data = datos)
##
## Coefficients:
## (Intercept)          X2T          X1T
##      0.5923       0.4764       0.4504
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mod.simplificado)
##
##                     Value p-value                Decision
## Global Stat        5.3097  0.2570 Assumptions acceptable.
## Skewness           0.9802  0.3222 Assumptions acceptable.
## Kurtosis           0.7724  0.3795 Assumptions acceptable.
## Link Function      1.3460  0.2460 Assumptions acceptable.
## Heteroscedasticity 2.2112  0.1370 Assumptions acceptable.
```

Seguidamente se analizará cada supuesto por separado.

1. Normalidad:

En este punto representamos el llamado QQ plot. Vemos la tendencia lineal de los residuos indicativa de la aceptación de normalidad. Esta idea queda reforzada al aplicar el test de Shapiro-Wilk para normalidad.

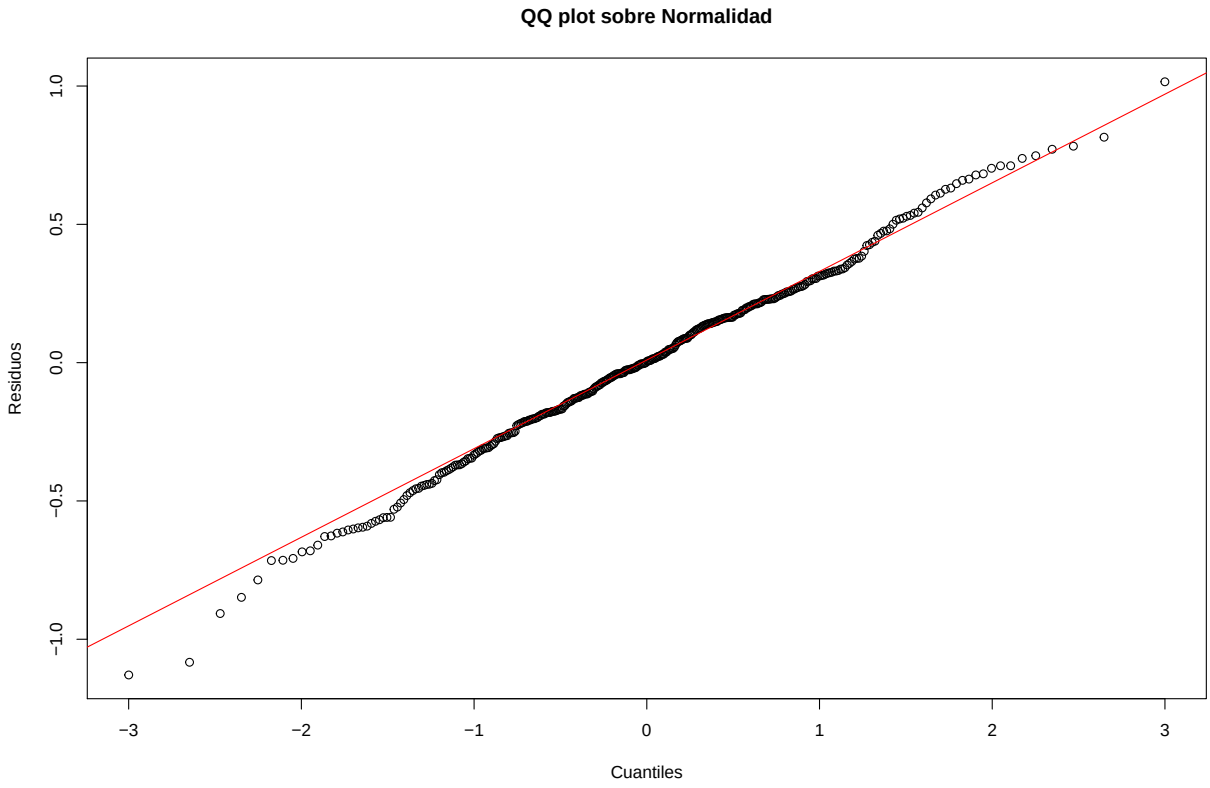


Figura 3.7: QQ plot sobre Normalidad

```
##
##  Shapiro-Wilk normality test
##
## data:  mod.simplificado$residuals
## W = 0.99589, p-value = 0.449
```

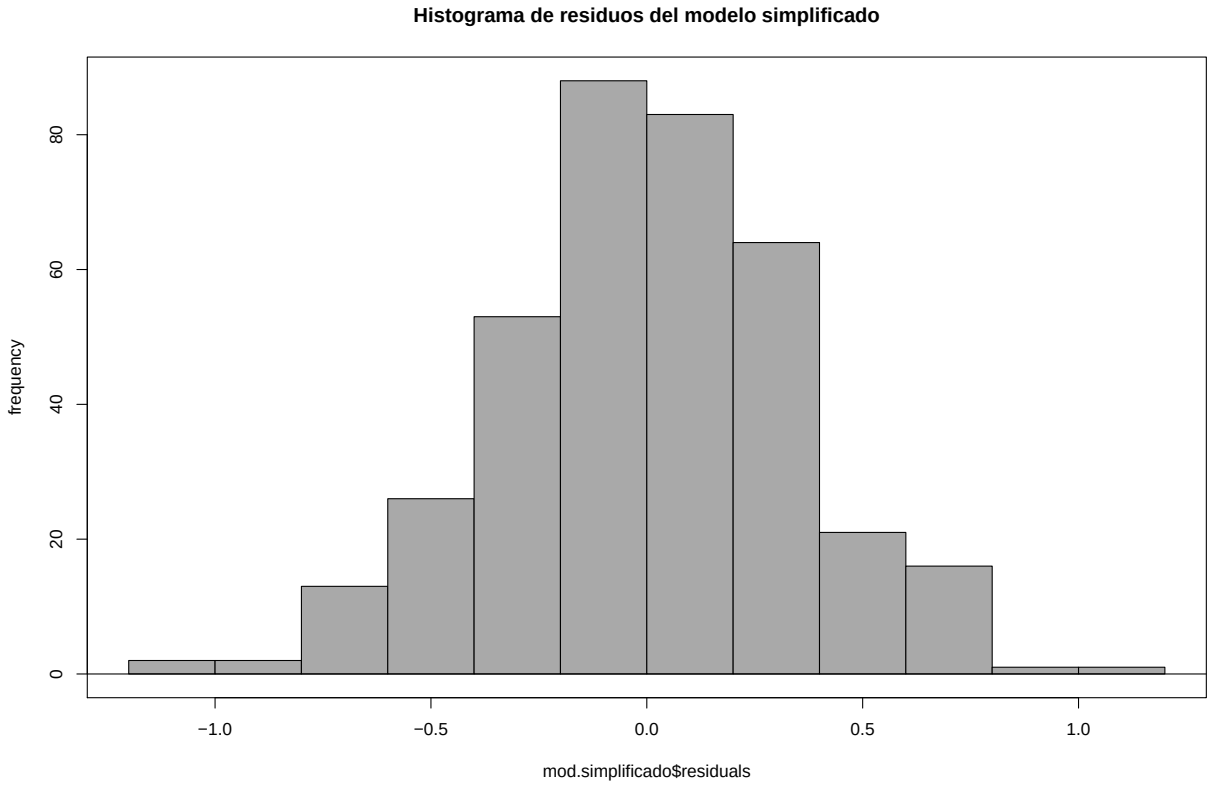


Figura 3.8: Histograma de residuos del modelo simplificado

2. Independencia:

Para el contrastar la hipótesis de independencia representamos los residuos según el orden en que fueron obtenidas las observaciones.

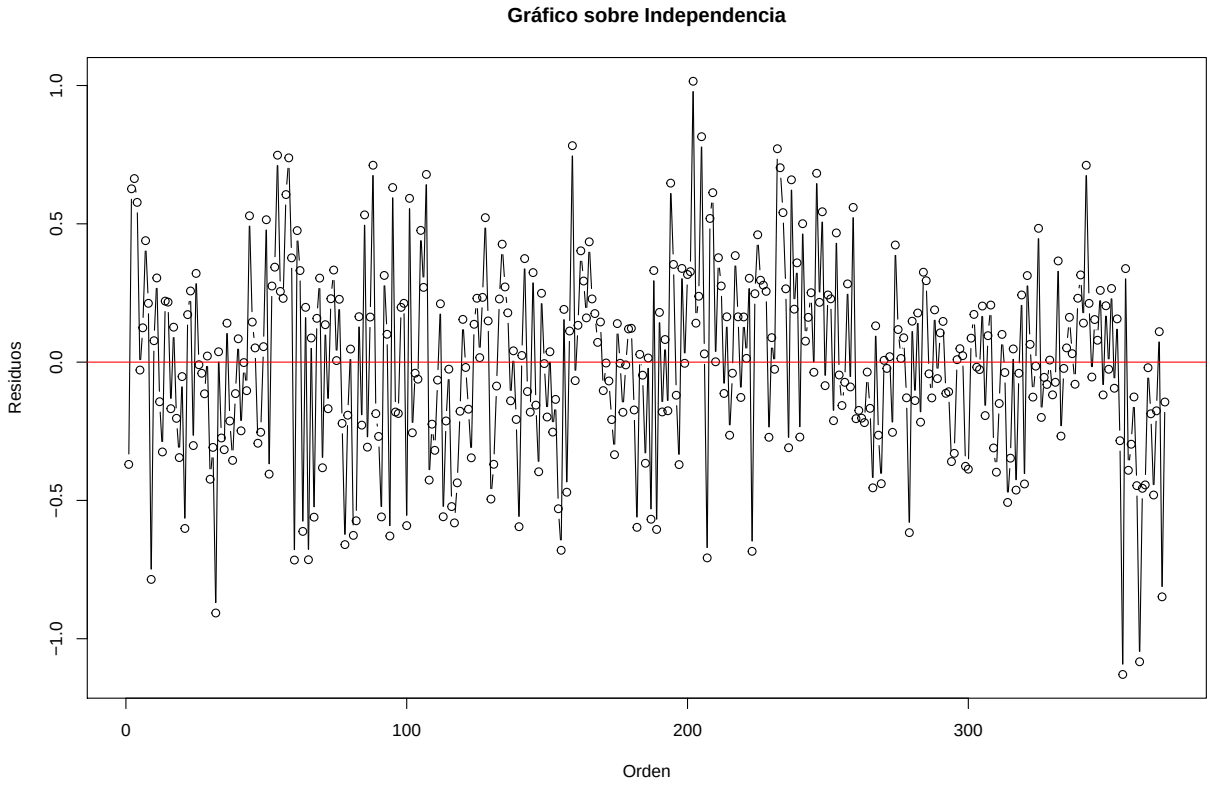


Figura 3.9: Gráfico sobre Independencia

El gráfico refleja la alternancia aleatoria de los puntos representados sin mostrar ningún patrón, indicativo claro de aceptación de la hipótesis de independencia.

3. Homocedasticidad:

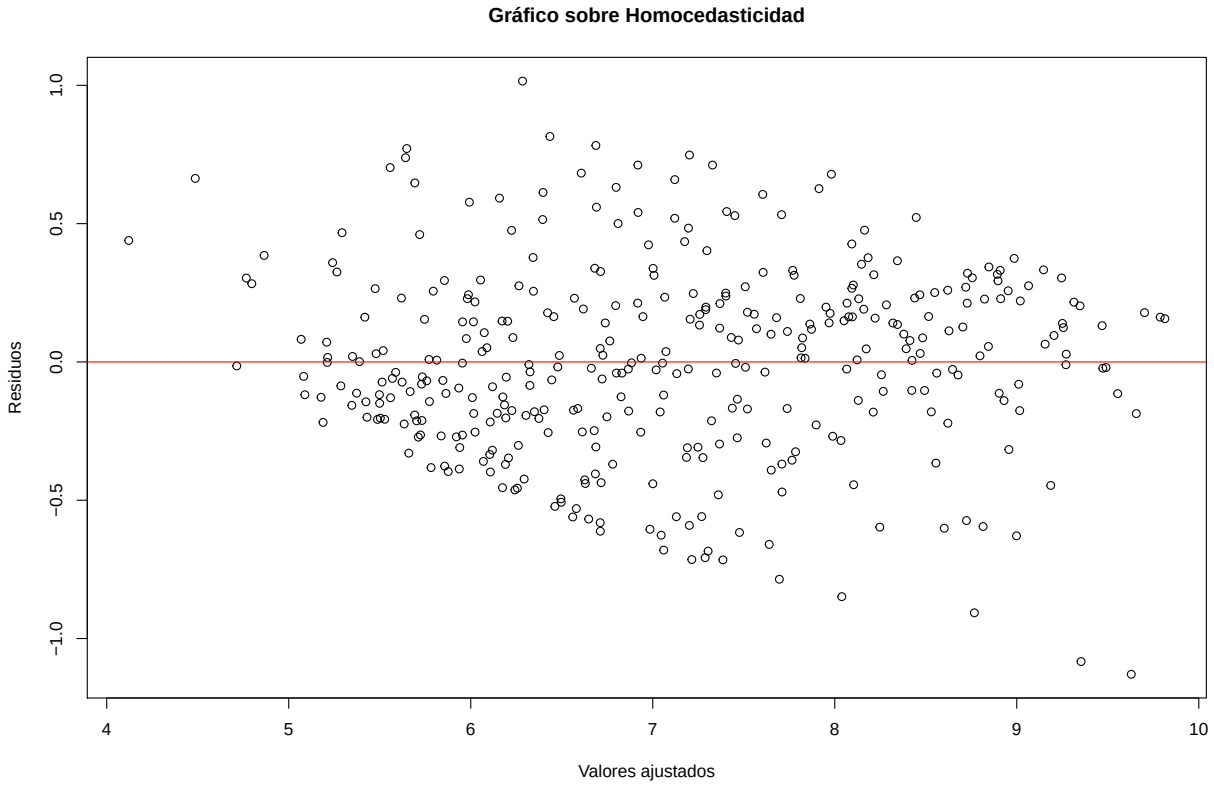


Figura 3.10: Gráfico sobre Homocedasticidad

Obervamos in intervalo máximo de variación de los residuos más o menos de la misma amplitud indicativo de la homocedasticidad.

3.2.4. Resultados y conclusiones

Aunque a priori puede parecer que el considerar el máximo de variables disponibles nos proporcionará un MRLM mejor que quitando alguna de ellas, debemos siempre valorar dos aspectos fundamentales:

- El sobreajuste por exceso, esto es, construir un modelo con muchas variables y obtener un buen ajuste para los datos con los que construimos dicho modelo no garantiza que para otra muestra distinta el ajuste sea igual de bueno. Es necesario analizar las correlaciones entre variables para seleccionar aquellas que efectivamente son candidatas para dicho fin.
- El coste computacional puede ser considerable si nuestro modelo se construye con muchas variables, un modelo más simple con resultados similares en cuanto al ajuste se refiere, puede ser en ocasiones la decisión más acertada, más aún si los valores de alguna/s variable/s aparecen como faltantes, outliers, etc.

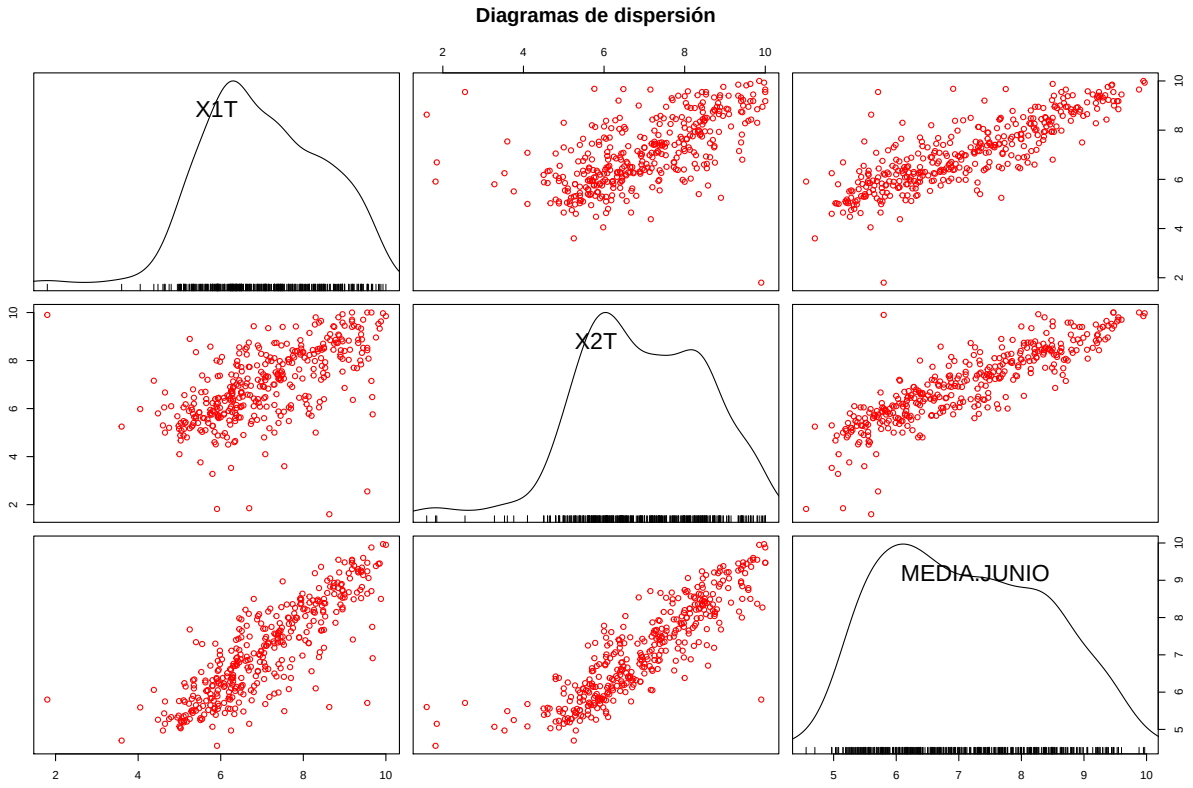


Figura 3.11: Diagramas de dispersión

Los resultados obtenidos con el modelo simplificado (sólo dos variables predictoras) han sido prácticamente idénticos a los obtenidos con el modelo completo (formado por siete variables predictoras). Si analizamos la figura anterior vemos que las distribuciones de las variables que componen el modelo simplificado es muy similar.

Para mostrar visualmente el ajuste realizado, se representan gráficamente los valores reales de las calificaciones (MEDIA.JUNIO) frente a los valores ajustados por el modelo:

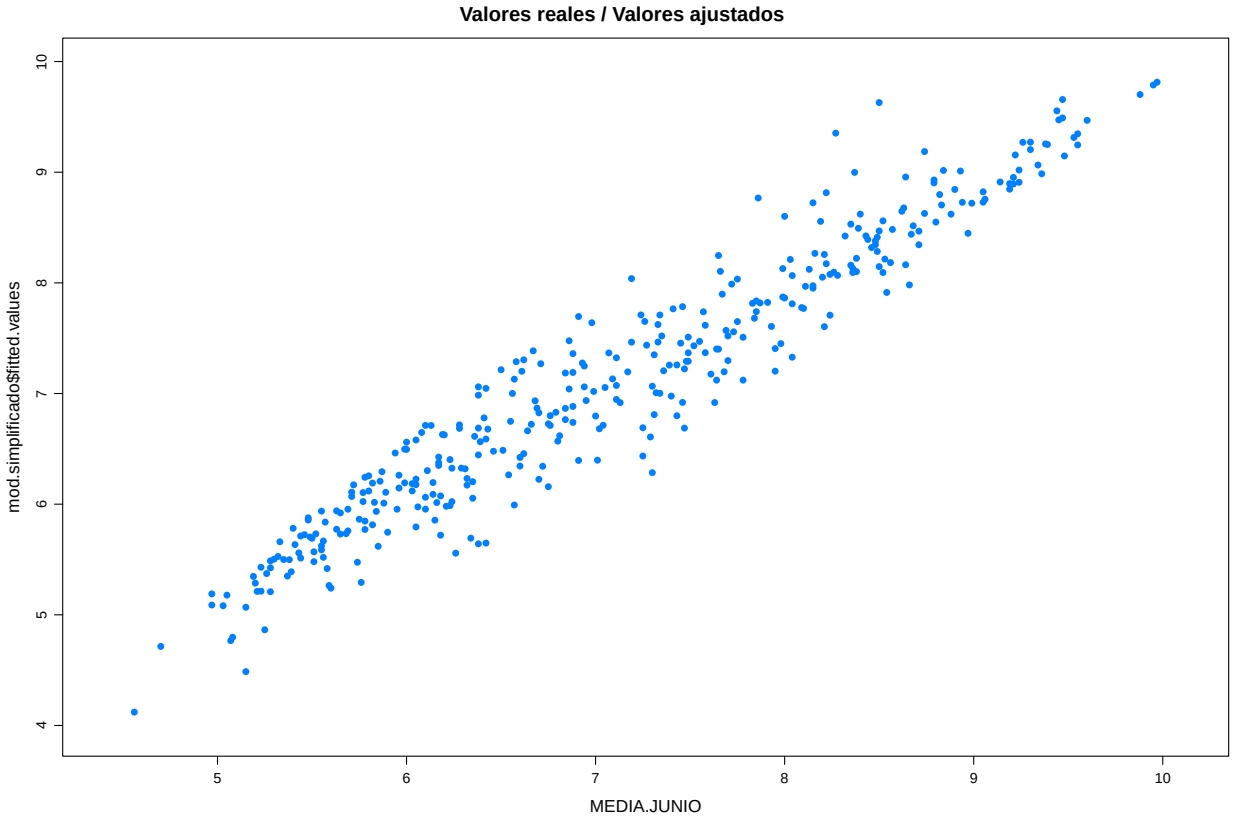


Figura 3.12: Valores reales / Valores ajustados

A modo de conclusión, podemos decir que el modelo propuesto ajusta bastante bien para los casos de alumnos aprobados (MEDIA.JUNIO >= 5). Para los suspensos es preciso seguir trabajando en la línea de intentar que sean más constantes en el trabajo durante cada trimestre.

Capítulo 4

Conclusiones, aportaciones y trabajos futuros

4.1. Conclusiones

Al comienzo del proyecto llevado a cabo, habían sido varios los intentos de conseguir procedimientos más o menos formales que ayudaran a estudiar la realidad de la actividad académica del centro. Hasta ese momento lo que se tenía era, por parte de cada unidad o departamento, un resumen personal de distintos aspectos que cada profesor había recabado manualmente según sus preferencias o intereses.

Con las distintas metodologías y técnicas usadas en este TFG, se ha conseguido dar la rigurosidad necesaria para que los resultados obtenidos tengan los fundamentos técnico-matemáticos que a todo trabajo científico se debe exigir.

4.2. Aportaciones

Fruto de la realización de este proyecto, el centro se ha enriquecido enormemente teniendo ahora varias herramientas para saber *qué está pasando* y *qué es probable que suceda* de cara a tomar decisiones de una forma más acertada que como se había hecho hasta la fecha. Concretamente, los aportes podemos resumirlos en:

- Sistema de **obtención de datos brutos** a partir de información visual mostrada en diferentes páginas web ofrecidas por la plataforma educativa Moodle.
- Diferentes programas informáticos construidos con el lenguaje de programación R para la depuración previa de los datos brutos, con el objetivo de detectar, corregir y transformar dichos datos brutos hasta **obtener conjuntos de datos válidos** para su tratamiento estadístico.

- Diferentes programas informáticos construidos con los lenguajes de programación R, RMarkdown, Lattex y HTML para la **generación automática de informes** en los formatos PDF y HTML atendiendo a la forma final de presentación.
- Implementación en R de protocolos de **metodologías para la construcción de distintos modelos de predicción** según las necesidades propias del centro.
- **Aportación al profesorado** de los distintos productos o resultados obtenidos, ofreciendo herramientas que ayuden a ampliar el conocimiento de la actividad diaria de sus aulas.

4.3. Trabajos futuros

Una vez concluido el curso académico y puesto en práctica la metodología y herramientas elaboradas, es lógico que se *"encienda la bombilla"* surgiendo propuestas de mejora y nuevas líneas de trabajo para ir ampliando y completando el denominado **sistema técnico de toma de decisiones**. En este sentido, algunas de las *asignaturas pendientes* a desarrollar son:

- Clasificar al alumnado según distintas categorías de actividad, es decir, entrega al menos una tarea, se presenta al menos a una prueba presencial, constante en el trabajo durante dos trimestres, etc.
- Estadísticas de accesos a cada uno de los enlaces incluidos en un aula, comunmente llamados *visitas* o *clics* para conocer mejor las preferencias del alumnado, secciones de mayor interés, detección de información no consultada, etc.
- Estadísticas descriptivas por bloques homogéneos, por ejemplo, asignaturas de la rama científica (matemáticas, física, etc) para afinar más en las predicciones de los modelos y obtener conclusiones más específicas, separadas de la tendencia general.
- Sistema automatizado de informes sobre la realización de tareas y cuestionarios online: tiempo de realización, tiempo de permanencia en la plataforma, días de la semana y horas más frecuentes de visita, etc.

Capítulo 5

Anexos

Los anexos que se especifican a continuación constituyen una serie de scripts en el lenguaje de programación R. El objetivo de cada uno de ellos es aplicar las técnicas expuestas en los capítulos anteriores para resolver computacionalmente los problemas, trabajos que se han ido describiendo a lo largo de este TFG fruto de las necesidades propias del centro educativo al que se hace referencia.

En dichos scripts sólo se hace hincapié en la aplicación y resolución de la parte técnica, usando para el tema de la presentación visual, e incluso para diseño y realización de este documento, los lenguajes RMarkdown, Latex y HTML.

5.1. Anexo 1: Informe automático de estadísticas descriptivas

El *objetivo* de este script es obtener una serie de estadísticas descriptivas de un trimestre para un curso y enseanza concreta. Se parte como *entrada* de dos ficheros CSV de la enseñanza considerada (uno por cada curso) y se obtiene como *salida* un informe en formato PDF o HTML (según se necesite) y un archivo CSV con los resultados.

```
#######################################################################
# TFG: Estadísticas Descriptivas
#
#       Objetivo del análisis: Un trimestre
#
#       Entrada:  trimestre, informe1.csv, informe2.csv, aulas.csv
#
#       Salida:   informe.csv
#######################################################################

#######################################################################
```

```
# Parámetros iniciales:

#cat("Entrada de datos:\n")
#cat("Número de trimestre: ") ; trimestre<-
#                                  as.numeric(scan("",what=character(),1))
objetivo<- "Un_trimestre" #Qué se analiza.
                          #Se usará en el nombre del archivo de resultados
trimestre<- 2 #Trimestre que queremos analizar

#########################################################################

# Funciones:

cambiaNaNporCero<- function(valores){
  resu<- valores
  for(i in 1:nrow(valores)){
    for(j in 2:ncol(valores)){
      resu[i,j]<- replace(resu[i,j],is.na(resu[i,j]),0)
    }
  }
  return(resu)
}

#Para todos los alumnos:
calcula<- function(valores,grupos,funcion,redondeo,nombre){
  resu <- by(valores,grupos,funcion)
  resu <- do.call(rbind,as.list(resu))
  resu<- data.frame(resu)
  resu<- round(resu,redondeo)
  resu<- cbind.data.frame(row.names(resu),resu)
  colnames(resu)<- c("Aula",nombre)
  resu<- na.omit(resu)
  row.names(resu)<- seq(1:nrow(resu))
  return(resu[2])
}

#Para alumnos activos:
calculaac<- function(curso,valores,columna,funcion,redondeo,nombre){
  resu <- valores
  resu <- tapply(resu[columna][,1],resu$Nomenclatura,funcion)
  resu <- do.call(rbind,as.list(resu))
  resu<- data.frame(resu)
  resu<- round(resu,redondeo)
  resu<- cbind.data.frame(row.names(resu),resu)
```

```
  colnames(resu)<- c("Aula",nombre)
  row.names(resu)<- seq(1:nrow(resu))
  resu<- replace(resu,is.na(resu),0)
  for(j in 2:ncol(resu)){
    resu[,j]<- replace(resu[,j],is.infinite(resu[,j]),0)
  }
  resu <- subset(resu, subset=(substr(resu$Aula,1,1)==curso))
  return(resu[2])
}

#########################################################################

# Carga de datos:

#fichero1<- file.choose() #Informe de Tutores de 1º curso a elegir
#fichero2<- file.choose() #Informe de Tutores de 2º curso a elegir
#aulas<- data.frame(read.csv(file.choose(),header=T,sep=";",dec=",",
                                              #fileEncoding="UTF-8"))
fichero1<- "TFG-4-inf_tutores_1_BTO.csv"
fichero2<- "TFG-5-inf_tutores_2_BTO.csv"
aulas<- data.frame(read.csv("TFG-6-aulasbach_sindatos.csv",header=T,sep=";",
                            dec=",",fileEncoding="UTF-8"))
fichero1.info<- file.info(fichero1)
fichero2.info<- file.info(fichero2)
fecha<- strftime(as.Date(fichero1.info$atime),
                 format="%Y_%m_%d") #Fecha de creación del fichero.
fechag<- strftime(as.Date(fichero1.info$atime),
                  format="%d/%m/%Y") #Fecha para gráficos.
datos1<- data.frame(read.csv(fichero1,header=T,sep=";",dec=",",
                             fileEncoding="UTF-8"))
datos2<- data.frame(read.csv(fichero2,header=T,sep=";",dec=",",
                             fileEncoding="UTF-8"))
datos<- rbind.data.frame(datos1,datos2)
datos$ID<- factor(datos$ID)
datos$NOTAS<- as.character(datos$NOTAS)

#Nombres válidos de columnas:
colnames(datos)[which(names(datos) == "Primer.acceso")] <-"Primer_acceso"
colnames(datos)[which(names(datos) == "Ultimo.acceso")] <-"Ultimo_acceso"
colnames(datos)[which(names(datos) == "TT.1T")] <-"TT_1T"
colnames(datos)[which(names(datos) == "TE.1T")] <-"TE_1T"
colnames(datos)[which(names(datos) == "TC.1T")] <-"TC_1T"
colnames(datos)[which(names(datos) == "TA.1T")] <-"TA_1T"
colnames(datos)[which(names(datos) == "NEP.1T")] <-"NEP_1T"
```

```
colnames(datos)[which(names(datos) == "NRJ.1T")] <-"NRJ_1T"
colnames(datos)[which(names(datos) == "NRS.1T")] <-"NRS_1T"
colnames(datos)[which(names(datos) == "TT.2T")] <-"TT_2T"
colnames(datos)[which(names(datos) == "TE.2T")] <-"TE_2T"
colnames(datos)[which(names(datos) == "TC.2T")] <-"TC_2T"
colnames(datos)[which(names(datos) == "TA.2T")] <-"TA_2T"
colnames(datos)[which(names(datos) == "NEP.2T")] <-"NEP_2T"
colnames(datos)[which(names(datos) == "NRJ.2T")] <-"NRJ_2T"
colnames(datos)[which(names(datos) == "NRS.2T")] <-"NRS_2T"
colnames(datos)[which(names(datos) == "TT.3T")] <-"TT_3T"
colnames(datos)[which(names(datos) == "TE.3T")] <-"TE_3T"
colnames(datos)[which(names(datos) == "TC.3T")] <-"TC_3T"
colnames(datos)[which(names(datos) == "TA.3T")] <-"TA_3T"
colnames(datos)[which(names(datos) == "NEP.3T")] <-"NEP_3T"
colnames(datos)[which(names(datos) == "NRJ.3T")] <-"NRJ_3T"
colnames(datos)[which(names(datos) == "NRS.3T")] <-"NRS_3T"
colnames(datos)[which(names(datos) == "FINAL.JUNIO")] <-"NOTA_JUN"
colnames(datos)[which(names(datos) == "FINAL.SEPTIEMBRE")] <-"NOTA_SEP"
colnames(datos)[which(names(datos) == "X1T")] <-"NOTA_1T"
colnames(datos)[which(names(datos) == "X2T")] <-"NOTA_2T"
colnames(datos)[which(names(datos) == "X3T")] <-"NOTA_3T"
colnames(datos)[which(names(datos) == "X1T.SENECA")] <-"NOTA_1T_SEN"
colnames(datos)[which(names(datos) == "X2T.SENECA")] <-"NOTA_2T_SEN"
colnames(datos)[which(names(datos) == "X3T.SENECA")] <-"NOTA_3T_SEN"

#Operaciones con fechas:
fecha_fichero<- strftime(as.Date(fichero1.info$atime),
                         format="%Y-%m-%d") #Fecha de creación del fichero.
datos$Primer_acceso<-as.Date(datos$Primer_acceso,format="%d/%m/%Y")
datos$Ultimo_acceso<-as.Date(datos$Ultimo_acceso,format="%d/%m/%Y")
datos$Dias_desde_primer_acceso<-
  with(datos,as.integer(difftime(fecha_fichero,datos$Primer_acceso,
                                 units="days")))
datos$Dias_desde_ultimo_acceso<-
  with(datos,as.integer(difftime(fecha_fichero,datos$Ultimo_acceso,
                                 units="days")))

#Limpiar datos no válidos (cuatrimestrales y pilotos):
datosbrutos<- datos
datos<- subset(datosbrutos, !(grepl("uatrimestre",Aula)))
datos<-
  subset(datos,
         !(grepl("1º Bach - Fundamentos del arte \\[Juan Luis Martínez\\]",
                      Aula)))
```

```
datos<-
  subset(datos,
         !(grepl("1º Bach - Historia del Mundo Contemporáneo \\
                 [Luis Rafael Villalta\\]",Aula)))
datos<-
  subset(datos,
         !(grepl("1º Bach - Lengua Castellana y Literatura \\
                 [Luis Miguel Flor, Nacho Vallejo\\]",Aula)))
row.names(datos)<- seq(1:nrow(datos))
n<- nrow(datos)
rm(datosbrutos)

#Obtención de nuevos campos:
for(i in 1:n){
  if(substr(datos$Aula[i],7,8)=="II"){
    datos$Curso[i]<- with(datos,"2")
  }else if (substr(datos$Aula[i],7,8)=="I "){
    datos$Curso[i]<- with(datos,"1")
  }else{
    datos$Curso[i]<- with(datos,substr(datos$Aula[i],1,1))
  }
}
for(i in 1:n){
  if(datos$TE_1T[i]>0){
    datos$AC_1T[i]<- with(datos,T)
  }else{
    datos$AC_1T[i]<- with(datos,F)
  }
}
Ense<- substr(datos$Aula,4,5)
for (i in 1:n){
  if (Ense[i]=="Ba"){
    datos$Nivel_educativo[i]<- with(datos,"Bachillerato")
  }else if (Ense[i]=="el"){
    datos$Nivel_educativo[i]<- with(datos,"ESA")
  }
}
Curso_Asignatura<- c("","")
for (i in 1:n){
  Curso_Asignatura<- strsplit(as.character(datos$Aula[i])," \\[")
}
Asignatura.Profesor<- ""
for (i in 1:n){
  Asignatura.Profesor[i]<- strsplit(as.character(datos$Aula)," - ")[[i]][2]
```

```
}
Asignatura_Profesor<- c("","")
for (i in 1:n){
  Asignatura_Profesor[i]<- strsplit(Asignatura.Profesor[[i]]," \\[")
}
Asignatura<- ""
for (i in 1:n){
  Asignatura[i]<- Asignatura_Profesor[[i]][1]
}
Profesor.corchete<- ""
for (i in 1:n){
  Profesor.corchete[i]<- Asignatura_Profesor[[i]][2]
}
Profesor<- ""
for (i in 1:n){
  Profesor[i]<- as.character(strsplit(Profesor.corchete[i],"\\]")[1])
}
datos$Asignatura<- with(datos,Asignatura)
datos$Profesor<- with(datos,Profesor)

#Eliminamos registros con aulas sin profesor:
for (i in 1:nrow(datos)){
  if (is.na(datos$Profesor[i])){
    datos<- datos[-i,]
  }
}
row.names(datos)<- seq(1:nrow(datos))
n<- nrow(datos)

#Ordenamos por apellidos:
datos<- datos[order(datos$Curso,datos$Apellidos),]
row.names(datos)<- seq(1:nrow(datos))

#Profes:
datos$Profes <- with(datos, NA)
profesores<- datos$Profesor
profes<- ""
for(i in 1:length(profesores)){
  profes<- ""
  if(length(grep(", ",profesores[i]))){
    p<- strsplit(profesores[i],", ")
    for(j in 1:length(p[[1]])){
      pro<- strsplit(p[[1]][j]," ")
      prok<- ""
```

```
      for (k in 1:length(pro[[1]])){
        prok<- paste(prok,substr(pro[[1]][k],1,1),sep="")
      }
      profes<- paste(profes,",",prok,sep="")
    }
    profes<- substr(profes,2,(nchar(profes)))
  }else{
    pro<- strsplit(profesores[i]," ")
    prok<- ""
    for (k in 1:length(pro[[1]])){
      prok<- paste(prok,substr(pro[[1]][k],1,1),sep="")
    }
    profes<- paste(profes,prok,sep="")
  }
  datos$Profes[i]<- profes
}

#Nomenclaturas:
datos$Nomenclatura <- with(datos, NA)
for (j in 1: (nrow(aulas))){
  for (i in 1:(nrow(datos))){
    if (as.character(datos$Aula[i])==as.character(aulas$Aula[j])){
      datos$Nomenclatura[i]<- paste(aulas$Curso[j],"-",
                                    aulas$Nomenclatura[j],
                                    "[",datos$Profes[i],"]",sep="")
    }
  }
}
datos$Nomenclatura<- factor(datos$Nomenclatura)
datos$Curso<- factor(datos$Curso)

#Datos por Nivel educativo:
niveleducativo<- datos$Nivel_educativo[1]
ne<- toupper(substr(datos$Nivel_educativo[1],1,3))
datos1 <- subset(datos,
                 subset=((Nivel_educativo==niveleducativo)&(Curso==1)))
datos2 <- subset(datos,
                 subset=((Nivel_educativo==niveleducativo)&(Curso==2)))
row.names(datos1)<- seq(1:nrow(datos1))
row.names(datos2)<- seq(1:nrow(datos2))
datos<- rbind.data.frame(datos1,datos2)


######################################################################
```

```
# Análisis:

ldatos<- list(datos1,datos2)
trinivel1<- data.frame() #Para nivel 1 (1º)
trinivel2<- data.frame() #Para nivel 2 (2º)

tricerrado_original<- function(j){
  estado<- 0 #Trimestre abierto
  if(j==1){
    if(max(datos$NOTA_1T_SEN)>0){
      estado<- 1
    }
  }else if(j==2){
    if(max(datos$NOTA_2T_SEN)>0){
      estado<- 1
    }
  }else if(j==3){
    if(max(datos$NOTA_3T_SEN)>0){
      estado<- 1
    }
  }else if(j==4){
    if(max(datos$NOTA_JUN)>0){
      estado<- 1
    }
  }else if(j==5){
    if(max(datos$NOTA_SEP)>0){
      estado<- 1
    }
  }
  return(estado)
}

tricerrado<- function(j){
  estado<- 0 #Trimestre abierto
  if(max(datos[paste("NOTA_",j,"T_SEN",sep="")])>0){
    estado<- 1
  }
  return(estado)
}

estadCerradas<- function(i,columnas,titulos){
  trim<- data.frame(sort(unique(ldatos[[i]]$Nomenclatura)))
  colnames(trim)<- "Aula"
  trim<- data.frame(trim,calcula(ldatos[[i]][columnas][1][,1],
```

```
                                ldatos[[i]]$Nomenclatura,mean,2,
                                titulos[1]))
trim<- data.frame(trim,calcula(ldatos[[i]][columnas][2][,1],
                                ldatos[[i]]$Nomenclatura,mean,2,
                                titulos[2]))
trim<-
  data.frame(trim,
             calculaac(i,
                       ldatos[[i]][which(ldatos[[i]][columnas][3][,1]>0),
                                         c("Nomenclatura",columnas[3],
                                           columnas[2])],columnas[2],
                                mean,2,titulos[3]))
trim<- data.frame(trim,calcula(ldatos[[i]][columnas][3][,1],
                                ldatos[[i]]$Nomenclatura,sum,0,
                                titulos[4]))
trim<- data.frame(trim,calcula(ldatos[[i]][columnas][4][,1],
                                ldatos[[i]]$Nomenclatura,sum,0,
                                titulos[5]))
trim<-
  data.frame(trim,calcula(ldatos[[i]][columnas][5][,1],
                          ldatos[[i]]$Nomenclatura,length,0,titulos[6]))
trim<-
  data.frame(trim,
             calculaac(i,
                       ldatos[[i]][which(ldatos[[i]][columnas][3][,1]>0),
                                           c("Nomenclatura",columnas[3],
                                           columnas[5])],columnas[5],
                                           length,2,titulos[7]))
trim<- data.frame(trim,
                  (round((trim[titulos[7]][,1]*100/
                          trim[titulos[6]][,1]),2)))
colnames(trim)[length(names(trim))]<- titulos[8]
trim<- data.frame(trim,
                  (round((trim[titulos[4]][,1]/
                          trim[titulos[7]][,1]),2)))
colnames(trim)[length(names(trim))]<- titulos[9]
trim<- data.frame(trim,(round((trim[titulos[4]][,1]*100/
                               calcula(ldatos[[i]][columnas][6][,1],
                                       ldatos[[i]]$Nomenclatura,
                                       sum,0,titulos[4])),2)))
colnames(trim)[length(names(trim))]<- titulos[10]
trim<- data.frame(trim,
                  calculaac(i,ldatos[[i]]
                            [which(ldatos[[i]][columnas][7][,1]!="NP"&
```

```
                                        ldatos[[i]][columnas][7][,1]!="SC"),
                                c("Nomenclatura",columnas[7],columnas[7])],
                                columnas[7],length,0,titulos[11]))
  trim<- data.frame(trim,
                    calculaac(i,ldatos[[i]]
                                [which(ldatos[[i]][columnas][7][,1]=="Apto"),
                                c("Nomenclatura",columnas[7],columnas[7])],
                                columnas[7],length,0,titulos[12]))
  trim<- data.frame(trim,
                    calculaac(i,ldatos[[i]]
                                [which(ldatos[[i]][columnas][1][,1]>=5),
                                c("Nomenclatura",columnas[1],columnas[1])],
                                columnas[1],length,0,titulos[13]))
  trim<- data.frame(trim,(round((trim[,14]*100/trim[,7]),2)))
  colnames(trim)[length(names(trim))]<- titulos[14]
  trim<- data.frame(trim,(round((trim[,14]*100/trim[,8]),2)))
  colnames(trim)[length(names(trim))]<- titulos[15]
  trim<- data.frame(trim,
                    calcula(ldatos[[i]][columnas][8][,1],
                            ldatos[[i]]$Nomenclatura,sum,0,titulos[16]))
  trim<- data.frame(trim,(trim[,17]/trim[,7]))
  colnames(trim)[length(names(trim))]<- titulos[17]
  trim<- data.frame(trim,(round((trim[,5]*100/trim[,17]),2)))
  colnames(trim)[length(names(trim))]<- titulos[18]
  trim<- data.frame(trim,(round((trim[,13]*100/trim[,12]),2)))
  colnames(trim)[length(names(trim))]<- titulos[19]
  trim<- data.frame(trim,(round((trim[,13]*100/trim[,7]),2)))
  colnames(trim)[length(names(trim))]<- titulos[20]
  trim<- data.frame(trim,(round((trim[,12]*100/trim[,7]),2)))
  colnames(trim)[length(names(trim))]<- titulos[21]
  trim<- cambiaNaNporCero(trim)
  return(trim)
}

estadAbiertas<- function(i,columnas,titulos){
  trim<- data.frame(sort(unique(ldatos[[i]]$Nomenclatura)))
  colnames(trim)<- "Aula"
  trim<- data.frame(trim,
                    calcula(ldatos[[i]][columnas][1][,1],
                            ldatos[[i]]$Nomenclatura,mean,2,titulos[1]))
  trim<- data.frame(trim,
                    calcula(ldatos[[i]][columnas][2][,1],
                            ldatos[[i]]$Nomenclatura,mean,2,titulos[2]))
  trim<- data.frame(trim,
```

```
                  calculaac(i,ldatos[[i]]
                            [which(ldatos[[i]][columnas][3][,1]>0),
                            c("Nomenclatura",columnas[3],columnas[2])],
                            columnas[2],mean,2,titulos[3]))
trim<- data.frame(trim,calcula(ldatos[[i]][columnas][3][,1],
                               ldatos[[i]]$Nomenclatura,sum,0,
                               titulos[4]))
trim<- data.frame(trim,calcula(ldatos[[i]][columnas][4][,1],
                               ldatos[[i]]$Nomenclatura,sum,0,
                               titulos[5]))
trim<- data.frame(trim,calcula(ldatos[[i]][columnas][5][,1],
                               ldatos[[i]]$Nomenclatura,length,0,
                               titulos[6]))
trim<- data.frame(trim,
                  calculaac(i,ldatos[[i]]
                            [which(ldatos[[i]][columnas][3][,1]>0),
                            c("Nomenclatura",columnas[3],columnas[5])],
                            columnas[5],length,2,titulos[7]))
trim<- data.frame(trim,
                  (round((trim[titulos[7]][,1]*100/
                          trim[titulos[6]][,1]),2)))
colnames(trim)[length(names(trim))]<- titulos[8]
trim<- data.frame(trim,
                  (round((trim[titulos[4]][,1]/
                          trim[titulos[7]][,1]),2)))
colnames(trim)[length(names(trim))]<- titulos[9]
trim<- data.frame(trim,(round((trim[titulos[4]][,1]*100/
                               calcula(ldatos[[i]][columnas][6][,1],
                                       ldatos[[i]]$Nomenclatura,
                                       sum,0,titulos[4])),2)))
colnames(trim)[length(names(trim))]<- titulos[10]
trim<-
  data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[11]
trim<-
  data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[12]
trim<-
  data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[13]
trim<-
  data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[14]
trim<-
  data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[15]
trim<-
  data.frame(trim,
             calcula(ldatos[[i]][columnas][8][,1],
```

```
                        ldatos[[i]]$Nomenclatura,sum,0,titulos[16]))
  trim<- data.frame(trim,(trim[,17]/trim[,7]))
  colnames(trim)[length(names(trim))]<- titulos[17]
  trim<- data.frame(trim,(round((trim[,5]*100/trim[,17]),2)))
  colnames(trim)[length(names(trim))]<- titulos[18]
  trim<-
    data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[19]
  trim<-
    data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[20]
  trim<-
    data.frame(trim,0) ; colnames(trim)[length(names(trim))]<- titulos[21]
  trim<- cambiaNaNporCero(trim)
  return(trim)
}

titulos<- c("NotaMeSEN",
            "NotaMe",
            "NotaMeAc",
           "NumTE",
           "NumTA",          #5
           "NumAlu",
           "NumAluAc",
           "PorAluAc",
           "MeTEAc",
           "PorTEAluAc",     #10
           "NumAluPre",
           "NumAluApto",
           "NumAluApro",
           "PorAproTot",
           "PorAproAc",      #15
           "NumTT",
           "NumTTAlu",
           "PorTE",
           "PorAluAptoPre",
           "PorAluAptoTot",  #20
           "PorAluPre")

titulos.descripcion<- c("Nota media en Séneca",
                        "Nota media de las tareas",
                        "Nota media de las tareas del alumnado activo",
                        "Nº tareas entregadas",
                        "Nº tareas aprobadas",
                        "Nº total de alumnos",
                        "Nº alumnos activos",
```

```
                        "% alumnos activos",
                        "Media tareas entregadas por alumno activo",
                        "% tareas entregadas por alumno activo",
                        "Nº alumnos presentados a la prueba presencial",
                        "Nº alumnos aptos en la prueba presencial",
                        "Nº alumnos aprobados en el trimestre",
                        "% aprobados sobre el total de alumnos",
                        "% aprobados sobre activos",
                        "Nº total tareas a evaluar",
                        "Nº total tareas a realizar por alumno",
                        "% tareas entregadas respecto al total a evaluar",
                        "% alumnos aptos de los presentados",
                        "% alumnos aptos del total de alumnos",
                        "% alumnos presentados")

#Para tablas y gráficos:
variables<- data.frame(titulos,titulos.descripcion)
colnames(variables)<- c("Variable","Descripción")

#Creación de tablas de resultados:
for(i in 1:2){ #Para cada curso (1:2)
  j<- trimestre
  if(tricerrado(j)==1){ #Trimestre cerrado
    assign(paste("trinivel",i,sep=""),
           estadCerradas(i,c(paste("NOTA_",j,"T_SEN",sep=""),
                             paste("NOTA_",j,"T",sep=""),
                             paste("TE_",j,"T",sep=""),
                             paste("TA_",j,"T",sep=""),
                             paste("DNI",sep=""),
                             paste("TT_",j,"T",sep=""),
                             paste("NEP_",j,"T",sep=""),
                             paste("TT_",j,"T",sep="")),
                         titulos),
           envir = .GlobalEnv)
  }else{
    assign(paste("trinivel",i,sep=""),
           estadAbiertas(i,c(paste("NOTA_",j,"T_SEN",sep=""),
                             paste("NOTA_",j,"T",sep=""),
                             paste("TE_",j,"T",sep=""),
                             paste("TA_",j,"T",sep=""),
                             paste("DNI",sep=""),
                             paste("TT_",j,"T",sep=""),
                             paste("NEP_",j,"T",sep=""),
                             paste("TT_",j,"T",sep="")),
```

```
                          titulos),
           envir = .GlobalEnv)
  }
}
tri<- list(trinivel1,trinivel2)
for(i in 1:2){
  assign(paste("trinivel",i,sep=""),
         replace(get(paste("trinivel",i,sep="")),
                 is.na(get(paste("trinivel",i,sep=""))),0),
         envir = .GlobalEnv)
}

#Aulas por curso:
aulas1<- data.frame(sort(unique(datos1$Aula)))
colnames(aulas1)<- c("Aula")
aulas2<- data.frame(sort(unique(datos2$Aula)))
colnames(aulas2)<- c("Aula")
```

El objetivo de este script es construir funciones que se usan en el script anterior, para hacer distintas representaciones gráficas según sean necesarias.

```
#######################################################################
# TFG: Estadísticas Descriptivas
#
#     Objetivo: Representaciones gráficas auxiliares
#######################################################################


#######################################################################
# Funciones:

graficop.param<- function(){
  par(las=2,cex.axis=0.78,mar=c(15,4,4,2)+0.1,lend=2,col.main="darkgreen",
      font.main=1)
}

graficop<- function(valores,columna,color,titulo,fechag){
  graficop.param()
  plot(valores[,columna],type="h",axes=FALSE,ann=FALSE,lwd=10,col=color)
  axis(1,at=1:(nrow(valores)),lab=valores[,1])
  axis(2,las=1,at=0:(max(valores[,columna],na.rm=TRUE)+1))
  title(main=paste(titulo,fechag))
  box()
  text(valores,labels=valores[,columna],adj=c(0.5,-0.8),pos=3,cex=0.7,
       col="darkgreen")
}

graficobp.param<- function(){
  #par(las=1,cex.axis=0.80,mar=c(4,15,4,2)+0.1,lend=2,col.main="black",
  font.main=1)
  par(las=1,cex.axis=1.2,mar=c(4,10,4,2)+0.1,lend=2,col.main="black",
      font.main=2,
      cex.main=2)
}

graficobp<- function(valores,columna,titulo,fechag){
  graficobp.param()
  bp<-barplot(valores[,columna], horiz=T, col=rainbow(nrow(valores)),
              xlim=c(0,max(valores[,columna])+2),main=paste(titulo,fechag),
              names.arg =valores[,1])
  text(valores[,columna],bp,labels=valores[,columna],pos=4,cex=0.9,
       col="black")
}
```

```
graficobpv.param<- function(){
  par(las=2,cex.axis=1.2,mar=c(15,4,4,2)+0.1,lend=2,col.main="black",
      font.main=1)
}

graficobpv<- function(valores,columna,titulo,fechag){
  graficobpv.param()
  bp<-barplot(valores[,columna], horiz=F, col=rainbow(nrow(valores)),
              xlim=c(0,max(valores[,columna])+1),main=paste(titulo,fechag),
              names.arg =valores[,1])
  text(valores[,columna],bp,labels=valores[,columna],pos=4,cex=0.7,
       col="#424242")
}

#Para comparar:
gra.bp.com.param_bak<- function(){
  par(las=1,cex.axis=1.2,mar=c(4,15,4,2)+0.1,lend=2,col.main="black",
      font.main=2,
      cex.main=2)
}

#Para comparar:
gra.bp.com_bak<- function(valores,categorias,titulo,fechag){
  gra.bp.com.param()
  bp<-barplot(as.matrix(t.data.frame(valores)), horiz=T,
              col=rainbow(ncol(valores)),
              xlim=c(0,max(valores)+1),main=paste(titulo,fechag),
              names.arg =categorias,beside=T,legend.text=colnames(valores))
  texto<-NA
  for(i in 1:nrow(valores)){
    for(j in 1:ncol(valores)){
      texto<- c(texto,valores[,j][i])
    }
  }
  texto<- texto[-1]
  text(texto,bp,labels=texto,pos=4,cex=0.9,col="black")
  #text(c(valores[,1],valores[,2]),bp,
  #labels=c(valores[,1],valores[,2]),pos=4,
  #cex=0.5,col="black")
}

#Para comparar:
gra.bp.com.param<- function(){
  par(las=1,cex.axis=1.2,mar=c(4,15,4,2)+0.1,lend=2,
```

```
      col.main="black",font.main=2,cex.main=2)
}

#Para comparar:
gra.bp.com<- function(valores,categorias,titulo,fechag){
  gra.bp.com.param()
  bp<-barplot(as.matrix(t.data.frame(valores)), horiz=T,
              col=rainbow(ncol(valores)),xlim=c(0,max(valores)+1),
              main=paste(titulo,fechag),names.arg=categorias,
              beside=F,legend.text=colnames(valores))
  texto<-NA
  for(i in 1:nrow(valores)){
    for(j in 1:ncol(valores)){
      texto<- c(texto,valores[,j][i])
    }
  }
  texto<- texto[-1]
  text(texto,bp,labels=texto,pos=4,cex=0.9,col="black")
  #text(c(valores[,1],valores[,2]),bp,
  #labels=c(valores[,1],valores[,2]),pos=4,
  #cex=0.5,col="black")
}
```

5.2. Anexo 2: Modelo de clasificación SVM para predicción de bajas de matrícula

El *objetivo* de este script es construir un modelo de predicción, usando la técnica de máquinas de vectores soporte (SVM), para clasificar los alumnos que pueden causar baja en su matrícula con un mes de antelación. Se parte como *entrada* de dos ficheros CSV: uno con los datos de febrero (antes de producirse las bajas) y otro con los datos de marzo (una vez producidas las bajas). Se obtienen como *salida* los resultados a modo de tabla de confusión, un gráfico que muestra las regiones de clasificación y un archivo CSV con los resultados relativos a las predicciones.

```
#######################################################################
# TFG: modelo.svms
#
#       Objetivo del modelo.svm: Bajas
#
#       Entrada:  informe_despues.csv, informe_antes.csv, aulas.csv
#
#       Salida:   informe_con_bajas.csv
#######################################################################

#######################################################################

# Parámetros iniciales:

objetivo<- "Bajas"  #Qué se analiza. Se usará en el nombre del archivo
                    #de resultados

#######################################################################

#fichero1<- file.choose() #Informe de Tutores DESPUÉS de bajas para elegir
#fichero2<- file.choose() #Informe de Tutores ANTES de bajas para elegir
#aulas<- data.frame(read.csv(file.choose(),header=T,sep=";",dec=",",
                            #fileEncoding="UTF-8"))
fichero1<- "TFG-1-inf_tutores_1_BTO_despues.csv"
fichero2<- "TFG-2-inf_tutores_1_BTO_antes.csv"
aulas<- data.frame(read.csv("TFG-3-aulasbach.csv",header=T,sep=";",dec=",",
                            fileEncoding="UTF-8"))
fichero1.info<- file.info(fichero1)
fichero2.info<- file.info(fichero2)
fecha<- strftime(as.Date(fichero1.info$mtime),
                 format="%Y-%m-%d") #Fecha de creación del fichero.
fechag<- strftime(as.Date(fichero1.info$mtime),
                  format="%d/%m/%Y") #Fecha para gráficos.
```

```
#aulasbach<- data.frame(read.csv("aulasbach.csv",header=T,sep=";",dec=",",
                                 #fileEncoding="UTF-8"))
datos1<- data.frame(read.csv(fichero1,header=T,sep=";",dec=","))
datos1$ID<- factor(datos1$ID)
datos1$NOTAS<- as.character(datos1$NOTAS)

datos2<- data.frame(read.csv(fichero2,header=T,sep=";",dec=","))
datos2$ID<- factor(datos2$ID)
datos2$NOTAS<- as.character(datos2$NOTAS)

#Nombres válidos de columnas:
colnames(datos1)[which(names(datos1) == "Primer.acceso")] <-"Primer_acceso"
colnames(datos1)[which(names(datos1) == "Ultimo.acceso")] <-"Ultimo_acceso"
colnames(datos1)[which(names(datos1) == "TT.1T")] <-"TT_1T"
colnames(datos1)[which(names(datos1) == "TE.1T")] <-"TE_1T"
colnames(datos1)[which(names(datos1) == "TC.1T")] <-"TC_1T"
colnames(datos1)[which(names(datos1) == "TA.1T")] <-"TA_1T"
colnames(datos1)[which(names(datos1) == "NEP.1T")] <-"NEP_1T"
colnames(datos1)[which(names(datos1) == "NRJ.1T")] <-"NRJ_1T"
colnames(datos1)[which(names(datos1) == "NRS.1T")] <-"NRS_1T"
colnames(datos1)[which(names(datos1) == "TT.2T")] <-"TT_2T"
colnames(datos1)[which(names(datos1) == "TE.2T")] <-"TE_2T"
colnames(datos1)[which(names(datos1) == "TC.2T")] <-"TC_2T"
colnames(datos1)[which(names(datos1) == "TA.2T")] <-"TA_2T"
colnames(datos1)[which(names(datos1) == "NEP.2T")] <-"NEP_2T"
colnames(datos1)[which(names(datos1) == "NRJ.2T")] <-"NRJ_2T"
colnames(datos1)[which(names(datos1) == "NRS.2T")] <-"NRS_2T"
colnames(datos1)[which(names(datos1) == "TT.3T")] <-"TT_3T"
colnames(datos1)[which(names(datos1) == "TE.3T")] <-"TE_3T"
colnames(datos1)[which(names(datos1) == "TC.3T")] <-"TC_3T"
colnames(datos1)[which(names(datos1) == "TA.3T")] <-"TA_3T"
colnames(datos1)[which(names(datos1) == "NEP.3T")] <-"NEP_3T"
colnames(datos1)[which(names(datos1) == "NRJ.3T")] <-"NRJ_3T"
colnames(datos1)[which(names(datos1) == "NRS.3T")] <-"NRS_3T"
colnames(datos1)[which(names(datos1) == "FINAL.JUNIO")] <-"NOTA_JUN"
colnames(datos1)[which(names(datos1) == "FINAL.SEPTIEMBRE")] <-"NOTA_SEP"
colnames(datos1)[which(names(datos1) == "X1T")] <-"NOTA_1T"
colnames(datos1)[which(names(datos1) == "X2T")] <-"NOTA_2T"
colnames(datos1)[which(names(datos1) == "X3T")] <-"NOTA_3T"
colnames(datos1)[which(names(datos1) == "X1T.SENECA")] <-"NOTA_1T_SEN"
colnames(datos1)[which(names(datos1) == "X2T.SENECA")] <-"NOTA_2T_SEN"
colnames(datos1)[which(names(datos1) == "X3T.SENECA")] <-"NOTA_3T_SEN"

colnames(datos2)[which(names(datos2) == "Primer.acceso")] <-"Primer_acceso"
```

```
colnames(datos2)[which(names(datos2) == "Ultimo.acceso")] <-"Ultimo_acceso"
colnames(datos2)[which(names(datos2) == "TT.1T")] <-"TT_1T"
colnames(datos2)[which(names(datos2) == "TE.1T")] <-"TE_1T"
colnames(datos2)[which(names(datos2) == "TC.1T")] <-"TC_1T"
colnames(datos2)[which(names(datos2) == "TA.1T")] <-"TA_1T"
colnames(datos2)[which(names(datos2) == "NEP.1T")] <-"NEP_1T"
colnames(datos2)[which(names(datos2) == "NRJ.1T")] <-"NRJ_1T"
colnames(datos2)[which(names(datos2) == "NRS.1T")] <-"NRS_1T"
colnames(datos2)[which(names(datos2) == "TT.2T")] <-"TT_2T"
colnames(datos2)[which(names(datos2) == "TE.2T")] <-"TE_2T"
colnames(datos2)[which(names(datos2) == "TC.2T")] <-"TC_2T"
colnames(datos2)[which(names(datos2) == "TA.2T")] <-"TA_2T"
colnames(datos2)[which(names(datos2) == "NEP.2T")] <-"NEP_2T"
colnames(datos2)[which(names(datos2) == "NRJ.2T")] <-"NRJ_2T"
colnames(datos2)[which(names(datos2) == "NRS.2T")] <-"NRS_2T"
colnames(datos2)[which(names(datos2) == "TT.3T")] <-"TT_3T"
colnames(datos2)[which(names(datos2) == "TE.3T")] <-"TE_3T"
colnames(datos2)[which(names(datos2) == "TC.3T")] <-"TC_3T"
colnames(datos2)[which(names(datos2) == "TA.3T")] <-"TA_3T"
colnames(datos2)[which(names(datos2) == "NEP.3T")] <-"NEP_3T"
colnames(datos2)[which(names(datos2) == "NRJ.3T")] <-"NRJ_3T"
colnames(datos2)[which(names(datos2) == "NRS.3T")] <-"NRS_3T"
colnames(datos2)[which(names(datos2) == "FINAL.JUNIO")] <-"NOTA_JUN"
colnames(datos2)[which(names(datos2) == "FINAL.SEPTIEMBRE")] <-"NOTA_SEP"
colnames(datos2)[which(names(datos2) == "X1T")] <-"NOTA_1T"
colnames(datos2)[which(names(datos2) == "X2T")] <-"NOTA_2T"
colnames(datos2)[which(names(datos2) == "X3T")] <-"NOTA_3T"
colnames(datos2)[which(names(datos2) == "X1T.SENECA")] <-"NOTA_1T_SEN"
colnames(datos2)[which(names(datos2) == "X2T.SENECA")] <-"NOTA_2T_SEN"
colnames(datos2)[which(names(datos2) == "X3T.SENECA")] <-"NOTA_3T_SEN"

#Operaciones con fechas:
datos1$Primer_acceso<-as.Date(datos1$Primer_acceso,format="%d/%m/%Y")
datos1$Ultimo_acceso<-as.Date(datos1$Ultimo_acceso,format="%d/%m/%Y")
fecha_inicio<- as.Date("2015-09-15",format="%Y-%m-%d")
fecha_fin<- as.Date(strftime(max(datos1$Ultimo_acceso,na.rm = TRUE)),
                    format="%Y-%m-%d")
#Quitamos los repetidoresque no entran:
datos1<- datos1[-which(datos1[,"Ultimo_acceso"]<fecha_inicio),]
for (i in 1:nrow(datos1)){ #Cambiamos Primer$acceso a fecha_inico a los
                           # repetidores que sí han entrado.
  if(!is.na(datos1$Primer_acceso[i])){
    if ((as.integer(difftime(datos1$Primer_acceso[i],fecha_inicio,
                             units="days")))<0){
```

```
      datos1$Primer_acceso[i]<- fecha_inicio
    }
  }
}
datos1$Dias_hasta_primer_acceso<-
  with(datos1,as.integer(difftime(datos1$Primer_acceso,fecha_inicio,
                                  units="days")))
datos1$Dias_desde_ultimo_acceso<-
  with(datos1,as.integer(difftime(fecha_fin,datos1$Ultimo_acceso,
                                  units="days")))

datos2$Primer_acceso<-as.Date(datos2$Primer_acceso,format="%d/%m/%Y")
datos2$Ultimo_acceso<-as.Date(datos2$Ultimo_acceso,format="%d/%m/%Y")
fecha_inicio<- as.Date("2015-09-15",format="%Y-%m-%d")
fecha_fin<- as.Date(strftime(max(datos2$Ultimo_acceso,na.rm = TRUE)),
                    format="%Y-%m-%d")

#Quitamos los repetidores que no entran:
datos2<- datos2[-which(datos2[,"Ultimo_acceso"]<fecha_inicio),]
#Cambiamos Primer$acceso a fecha_inico a los repetidores que sí han entrado:
for (i in 1:nrow(datos2)){
  if(!is.na(datos2$Primer_acceso[i])){
    if ((as.integer(difftime(datos2$Primer_acceso[i],fecha_inicio,
                             units="days")))<0){
      datos2$Primer_acceso[i]<- fecha_inicio
    }
  }
}
datos2$Dias_hasta_primer_acceso<-
  with(datos2,as.integer(difftime(datos2$Primer_acceso,fecha_inicio,
                                  units="days")))
datos2$Dias_desde_ultimo_acceso<-
  with(datos2,as.integer(difftime(fecha_fin,datos2$Ultimo_acceso,
                                  units="days")))

#Operaciones con fechas:
fecha_fichero1<- strftime(as.Date(fichero1.info$mtime),
                          format="%Y-%m-%d") #Fecha de creación del fichero.
datos1$Primer_acceso<-as.Date(datos1$Primer_acceso,format="%d/%m/%Y")
fecha_inicio<- "2015/09/15"
for (i in 1:nrow(datos1)){
  if(!is.na(datos1$Primer_acceso[i])){
    if ((as.integer(difftime(datos1$Primer_acceso[i],
                             fecha_inicio,units="days")))<0){
```

```
      datos1$Primer_acceso[i]<- "2015/09/15"
    }
  }
}
datos1$Ultimo_acceso<-as.Date(datos1$Ultimo_acceso,format="%d/%m/%Y")
datos1$Dias_desde_primer_acceso<-
  with(datos1,as.integer(difftime(fecha_fichero1,datos1$Primer_acceso,
                                  units="days")))
datos1$Dias_desde_ultimo_acceso<-
  with(datos1,as.integer(difftime(fecha_fichero1,datos1$Ultimo_acceso,
                                  units="days")))

fecha_fichero2<- strftime(as.Date(fichero2.info$mtime),
                          format="%Y-%m-%d") #Fecha de creación del fichero.
datos2$Primer_acceso<-as.Date(datos2$Primer_acceso,format="%d/%m/%Y")
for (i in 1:nrow(datos2)){
  if(!is.na(datos2$Primer_acceso[i])){
    if ((as.integer(difftime(datos2$Primer_acceso[i],
                             fecha_inicio,units="days")))<0){
      #datos1$Primer_acceso[i]<- strftime(as.Date(fecha_inicio),
      #                                   format="%d/%m/%Y")
      datos2$Primer_acceso[i]<- "2015/09/15"
    }
  }
}
datos2$Ultimo_acceso<-as.Date(datos2$Ultimo_acceso,format="%d/%m/%Y")
datos2$Dias_desde_primer_acceso<-
  with(datos2,as.integer(difftime(fecha_fichero2,datos2$Primer_acceso,
                                  units="days")))
datos2$Dias_desde_ultimo_acceso<-
  with(datos2,as.integer(difftime(fecha_fichero2,datos2$Ultimo_acceso,
                                  units="days")))

#Ordenamos por ID:
datos1<- datos1[order(datos1$ID),]
row.names(datos1)<- seq(1:nrow(datos1))

datos2<- datos2[order(datos2$ID),]
row.names(datos2)<- seq(1:nrow(datos2))

#Etiquetamos baja/no baja por alumno:
datos3<- datos2
datos3$Baja <- with(datos3, FALSE)
datos1porID<- data.frame(sort(unique(datos1$ID)))
```

```
colnames(datos1porID)<- c("ID")
datos2porID<- data.frame(sort(unique(datos2$ID)))
colnames(datos2porID)<- c("ID")
datos4<-data.frame(setdiff(datos2porID$ID,datos1porID$ID))
colnames(datos4)<- c("ID")
for (i in 1:nrow(datos4)){
  datos3$Baja[which(as.character(datos3$ID)==
                    as.character(datos4$ID[i]))]<-"SI"
}
for (i in 1:nrow(datos3)){
  if (datos3$Baja[i]==as.character("FALSE")){
    datos3$Baja[i]<- "NO"
  }
}
datos3$Baja<- factor(datos3$Baja)

#Etiquetamos a los activos:
datos3$Activo<- with(datos3,"NO")
datos3$Activo[which(datos3$TE_1T>0)]<- "SI"

write.table(datos3,file="InfTut1Bach_quienesbaja.csv",
            append=FALSE,quote=FALSE,sep=";",col.names=TRUE,
            qmethod=c("escape","double"),eol="\n",na="",dec=",",
            row.names=FALSE,fileEncoding= "")

### Clasificación por SVM ##########################################

#Conjunto de datos:
cor(datos3[,c("Dias_desde_ultimo_acceso","NOTA_1T")], use="complete")
Dias_desde_ultimo_acceso<- datos3["Dias_desde_ultimo_acceso"][,1]
NOTA_1T<- datos3["NOTA_1T"][,1]
Baja<- datos3["Baja"][,1]
datos<- data.frame(Dias_desde_ultimo_acceso,NOTA_1T,Baja)

#Gráfico sin modelo.svm
palette(c("darkblue","#F5A9A9"))
library(lattice)
xyplot(NOTA_1T~Dias_desde_ultimo_acceso,group=Baja,data=datos,cex=1,pch=13,
       main="Diagrama de dispersión",xlab="Días desde el último acceso",
       ylab="Nota 1º Trimestre",col=rainbow(2))

#Creamos una muestra para entrenar el modelo.svm
set.seed(12345)
elimina=sample(1:nrow(datos3),0.7*nrow(datos3))
```

```
muestra=datos[elimina,]
muestra<- na.omit(muestra)
entrena=datos[-elimina,]

#Construcción de modelo.svm con la librería e1071
library(e1071)
modelo.svm=svm(Baja~NOTA_1T+Dias_desde_ultimo_acceso,data=entrena,
               method="C-classification",
          kernel="radial",cost=10,gamma=.1,na.action=na.omit)

#Analizamos el comportamiento
predic=data.frame(predict(modelo.svm,muestra))
muestra=cbind(muestra,predic=predic)
colnames(muestra)<- c("Dias_desde_ultimo_acceso","NOTA_1T","Baja",
                      "Prediccion_Baja")
muestra[1:20,]
(confu<- table(muestra$Baja,muestra$Prediccion_Baja))
total<- confu[1,1]+confu[1,2]+confu[2,1]+confu[2,2]
bajas<- confu[2,1]+confu[2,2]
activos<- confu[1,1]+confu[1,2]

diasaviso<- min(muestra[which(muestra[,4]=="SI"),1])
notaaviso<- mean(muestra[which(muestra[,4]=="SI"),2])

#palette(c("darkblue","#FA5858"))
#plot(modelo.svm,data=datos,X1T.SENECA~Dias.desde.Ultimo.acceso,
#     symbolPalette = palette(rainbow(2)))
```

5.3. Anexo 3: Modelo de regresión lineal para estimación de calificaciones

El *objetivo* de este script es construir un modelo de predicción, usando la técnica de regresión lineal múltiple (MRLM), para estimar las calificaciones de la convocatoria ordinaria de mayo/junio, con los datos relativos a la primera y segunda evaluación. La *entrada* es un fichero CSV de un curso específico de una enseñanza. Se obtienen como *salida* los modelos completo (con todas las variables) y simplificado (reduciendo variables sin sacrificar el ajuste) y una evaluación gráfica y analítica del modelo simplificado considerado.

```
#######################################################################
# TFG:  modelo.rlm
#
#       Objetivo del modelo.rlm: Nota de Junio
#
#       Entrada:  TFG-informe_tutores_2BTO_mayo2016.csv
#######################################################################

#######################################################################

# Parámetros iniciales:

objetivo<- "Nota de Junio" #Qué se analiza.
                           #Se usará en el nombre del archivo de resultados

#######################################################################

datos<- data.frame(read.csv("TFG-informe_tutores_2BTO_mayo2016.csv",
                            header=T,sep=";",dec=",",fileEncoding="UTF-8"))

#Operaciones con fechas:
datos$Primer.acceso<-as.Date(datos$Primer.acceso,format="%d/%m/%Y")
datos$Ultimo.acceso<-as.Date(datos$Ultimo.acceso,format="%d/%m/%Y")
fecha_inicio<- as.Date("2015-09-15",format="%Y-%m-%d")
fecha_fin<- as.Date(strftime(max(datos$Ultimo.acceso,na.rm = TRUE)),
                    format="%Y-%m-%d")
datos$Dias.hasta.Primer.acceso<-
  with(datos,as.integer(difftime(datos$Primer.acceso,fecha_inicio,
                                 units="days")))
datos$Dias.desde.Ultimo.acceso<-
  with(datos,as.integer(difftime(fecha_fin,datos$Ultimo.acceso,
                                 units="days")))
datos$FINAL.JUNIO<- as.character(datos$FINAL.JUNIO)
datos<- datos[-which(datos$FINAL.JUNIO=="SC"),] #Posibles errores
```

```
datos$FINAL.JUNIO<- as.numeric(datos$FINAL.JUNIO)
datos<- datos[c("Primer.acceso","CE","CR","FPE","FDC","FPL","X1T",
                "X2T","X3T","FINAL.JUNIO")]
datos<- datos[-which(datos$X3T==0),]  #Los que dejan la asignatura
                                      #para septiembre.
datos<- datos[-which(datos$X3T<5),]   #Los que suspenden y van
                                      #a septiembre.
datos<- datos[-which(datos$X1T==0),]  #Los que dejan la asignatura
                                      #para septiembre.
datos<- datos[-which(datos$X2T==0),]  #Los que dejan la asignatura
                                      #para septiembre.
datos<- datos[-which(((datos$X1T)<5)&((datos$X2T)<5)),] #Los que van
                                                        #a Septiembre.
datos$MEDIA.JUNIO<- round(apply(datos[c("X1T","X2T","X3T")],1,mean),2)

datos$ERROR.JUNIO<- with(datos, (FINAL.JUNIO-MEDIA.JUNIO))

cor(datos[,c("CE","CR","FDC","FPE","FPL","MEDIA.JUNIO","X1T","X2T")],
    use="complete")

mod.completo <- lm(MEDIA.JUNIO~CE+CR+FDC+FPE+FPL+X1T+X2T, data=datos)
summary(mod.completo)

mod.simplificado <- lm(formula = MEDIA.JUNIO ~ X2T + X1T, data = datos)
summary(mod.simplificado)

#Test de normalidad:
shapiro.test(mod.simplificado$residuals)

#Test modelo adecuado:
library(gvlma)
gvlma(mod.simplificado)
```

Bibliografía

Alexander Borbón A., Walter Mora F. 2013. *Edición de textos científicos con lATEX*. Revista digital matemática. educación e internet. Instituto Tecnológico de Costa Rica.

Andrés Muñoz Ortega, Andrés Bueno Crespo y otros. s.f. «Big data: El valor añadido de los datos en su negocio». https://miriadax.net/web/big-data-el-valor-anadido-de-los-datos-en-su-negocio.

Álvaro Jiménez Galindo, Hugo Álvarez García. 2010. «Minería de datos en la educación». *Universidad Carlos III de Madrid* 1 (1): 1-8. doi:https://www.it.uc3m.es/jvillena/irc/practicas/10-11/08mem.pdf.

Blandón, Diego Alejandro Salazar. 2012. «Comparación de máquinas de soporte vectorial vs. regresión logística. ¿cuál es más recomendable para discriminar?» *Universidad Nacional de Colombia, Facultad de Ciencias, Escuela de Estadística Medellín, Colombia* 1 (1): 1-88. doi:http://www.bdigital.unal.edu.co/6594/1/15373359.2012.pdf.

Calot, Gérard. 1988. *Curso de estadística descriptiva*. Estadística. Paraninfo.

Crawley, Michael J. 2013. *The r book*. Estadística. Wiley.

Cuadras, C. M. 1991. *Métodos de análisis multivariante*. Estadística y análisis de datos. PPU.

David Masip Rodó, Jose Ramon Rodríguez Bermúdez y otros, Carles Garrigues Olivella. s.f. «Introducción al business intelligence y al big data (2.ª edición)». https://miriadax.net/web/estadistica-investigadores-3edicion.

M., Vargas Jiménez. 2008. «Ejemplo simple de regresión logística». *Departamento de Estadística e Investigación Operativa, Universidad de Granada* 1 (1): 8-9. doi:www.ugr.es/~mvargas/3.Ejesimpleregreslogistica.pdf.

Mejías, Rafael Pino. 2015. «Apuntes de clase de estadística computacional iI». *Dpto. de Estadística e Investigación Operativa, Universidad de Sevilla* 1 (1): 1-150.

RStudio. s.f. «R markdown: Dynamic documents for r». http://rmarkdown.rstudio.com/index.html.

Segundo, Fernando San. 2013. «Tutorial 10: Regresión». *Dpto. de Física y Matemáticas, Fac. de Biología, Uni. Alcalá de Henares* 1 (1): 16. doi:http://www3.uah.es/fsegundo/BioEstad/Tutorial-10.pdf.

Suárez, Enrique J. Carmona. 2014. «Tutorial sobre máquinas de vectores soporte (sVM)». *Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación*

aDistancia (UNED) 1 (1): 1-25. doi:http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf.

Venancio Tomeo Perucha, Isaías Uña Juárez. 2009. *Estadística descriptiva*. Ciencia estadística. Grupo Editorial Garceta.

Williams, Graham. 2011. *Data mining with rattle and r*. Use r! Springer.

Zoritza. 2008. «Ejemplo de regresión lineal múltiple». *Laboratorio docente de computación, Universidad Simón Bolívar, Facultad de Matemáticas y Sistemas* 1 (1): 1-15. doi:http://ldc.usb.ve/~moises/estadistica/Ej_Regresion_Lineal_Multiple_Zoritza.pdf.